LEARNING TO ADAPT UNDER PRACTICAL SENSING CONSTRAINTS

A Dissertation
Presented to
The Academic Faculty

by

Andrew Kenneth Massimino

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Electrical & Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2018

LEARNING TO ADAPT UNDER PRACTICAL SENSING CONSTRAINTS

Approved by:

Dr. Mark A. Davenport, Advisor
Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Yao Xie
Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Justin Romberg
Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Matthieu R. Bloch
Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Christopher J. Rozell
Electrical and Computer Engineering
Georgia Institute of Technology

Date approved: November 1st, 2018

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

The purpose of this work is to explore the capability of sensing systems to acquire information adaptively when they are subject to practical measurement constraints. By leveraging problem structure such as sparsity and probabilistic data models, intelligent sampling schemes have the potential to enable higher quality estimation with less sensing effort in diverse applications such as imaging, recommendation systems, information retrieval, and psychometric studies. Existing approaches to adaptive sensing are often limited in practice as they require the ability to take arbitrary measurements while in realistic situations, measurements must taken according to various limitations. Two representative constrained scenarios are considered: linear settings in which measurement rows are chosen from a fixed collection and where estimation may be performed only via sequentially chosen paired comparisons. Theoretical and empirical evidence are provided to suggest that adaptivity can result in substantial improvements in these constrained settings.

CHAPTER 1
INTRODUCTION

This thesis investigates the problem of acquiring signals via adaptively chosen measurements. Recent work in this area has shown that adaptive procedures can lead to significant improvement over non-adaptive ones in many circumstances. This is in contrast to standard results in compressed sensing, which suggest that any random selection of measurements should perform equally well—provided the collection of measurements to choose from satisfies certain strong *incoherence* conditions.

Unfortunately, existing approaches to adaptive sensing are limited in practice as they require the ability to take arbitrary linear measurements. In more realistic situations, measurements must be taken according to various practical *constraints*. The work presented in this thesis helps answer a number of questions, including determining which classes of measurements allow benefits under adaptive sensing and whether there are practical and efficient algorithms for choosing these measurements during the sensing process.

The answer to these questions will be highly dependent on appropriate *structured* signal models. For instance, we might consider a set of signals to be highly structured when its "size" (which can be defined in many ways) is much smaller than its ambient dimension and the set can be efficiently approached in a principled manner. As an example, the wavelet decomposition of natural images often has a tree structure—knowing any single coefficient tells you a great deal about other coefficients.

A well-developed theory of compressive sensing exists for the case where measurements are chosen non-adaptively and do not depend on prior observations. These results are attractive because they are signal agnostic and give uniform recovery guarantees (over the set of sparse signals). A major drawback to this approach is the large amount of sensing energy required by standard compressed sensing compared to a scenario in which the signal support were known and we could direct all of our effort to those coefficients. By design,

energy is spread over all of the ambient dimensions and all possible sparse supports.

It is reasonable to wonder whether recovery could be improved if we allow information collected during sensing to guide the acquisition process. Unfortunately, uniform guarantees for the performance under adaptivity are not substantially better than those of compressed sensing [1]. Essentially, noisy measurements of signals with very small components do not provide enough information about the signal's support to focus sensing energy onto the support quickly.

Nevertheless, several algorithms have recently been developed which have much better support recovery than compressed sensing does for sufficiently strong signals, see e.g., [2–4]. These results are obtained in an idealized situation where sensing vectors can be generated arbitrarily; it is not yet known whether more general adaptive sensing schemes can achieve this result in practical settings. Adaptivity does not help greatly over compressed sensing when: (i) the collection of measurements obeys the restricted isometry property (RIP) or (ii) signal components are too small. However, there are classes of problems where neither of these statements need be true. There is a vast unexplored middle-ground between the requirement of RIP ensembles and having full adaptive control over measurements. A few interesting practical situations where adaptivity is worth exploring are: magnetic resonance imaging (MRI) [5] using Fourier measurements of wavelet-tree-sparse signals; collection of radiation using a small number of active elements as in the single-pixel camera [6]; new analog-to-digital converters [7] taking highly quantized wideband signals; and psychometric studies [8] using human-generated data such as questionnaires.

The work in this thesis is an effort to help answer a number of questions; among these are determining which classes of measurements allow benefits under adaptive sensing and whether there are practical and efficient algorithms for choosing these measurements during the sensing process. Advancements in this area can lead to the development of efficient methods for adaptive sensing in constrained measurement settings and lead to powerful improvements in signal acquisition.

1.1 Contributions

The work in this thesis improves our understanding of sensing in two primary representative examples of constrained scenarios. The first is *constrained adaptive sensing*, where measurements must be drawn from a fixed collection which does not necessarily satisfy strong properties commonly assumed in compressed sensing. This topic is the focus of Chapter 2 and Chapter 5. In the second scenario, *localization via paired comparisons* we introduce a different type of constraint, where measurements are binary and are derived from distance information of items drawn in pairs. This scenario is studied in Chapters 3, 4 and 5.

Specifically, in Chapter 2, we propose a practical algorithm for constrained adaptive sensing by exploiting connections to optimal experimental design and show that these algorithms exhibit promising performance in some representative applications.

In Chapter 3, we prove theoretical bounds for how well we can expect to estimate a signal via paired comparisons under a randomized model of item generation in both the noiseless case and when the comparisons are noisy and subject to error. We show that random binary paired comparisons yield a stable embedding of the space of target signals and finally demonstrate that we can achieve significant gains by adaptively changing the item distribution.

In Chapter 4, we continue the study of paired comparisons, but where the items belong to a fixed embedding and no longer assumed to be Gaussian, greatly generalizing the previous pairwise comparison work. We discuss a fully sequential measurement approach guided by principles of information theory. We give bounds on the expected number of queries required to achieve a certain performance, and we validate our approach using simulated responses from a real-world dataset.

In Chapter 5, we revisit the measurement selection problem described first in Chapter 2. We tie the two problems constrained adaptive sensing and localization via paired comparisons together by showing a method of design which can work in both cases as well as apply to generalized linear models.

1.2 Background

> The power of adaption is critically dependent on our a-priori knowledge concerning the problem being studied; even a seemingly small change in the assumptions can lead to a different answer. — Erich Novak, *On the power of adaption [9].*

1.2.1 Compressed sensing

This thesis begins, in spirit, with the theory of compressive sensing, which we will sometimes abbreviate as (CS). By now, its basic tenet is quite well understood: if a quantity of interest which is nominally very high dimensional, has considerable structure, the amount of effort needed for sensing and estimation can be greatly reduced, as compared to classical methods [10–13]. As a brief overview, suppose we are trying to measure a signal $x \in \mathbb{R}^n$ where n is quite large. For a variety of reasons, e.g., to reduce cost, power consumption, or time, we may want to take merely m measurements where the number of measurements m is much smaller than n (this gives it the name *compressive* sensing). In many real-world problems, the measurement process can be abstractly modeled as a series of noisy linear observations;

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w} \iff y_i = \langle \mathbf{a}_i, \mathbf{x} \rangle + w_i, \ i = 1, \dots, m$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the "sensing matrix," $\mathbf{w} \in \mathbb{R}^m$ is a source of noise, and $\mathbf{y} \in \mathbb{R}^m$ is the vector of measurements. Using standard intuition from linear algebra, $\mathbf{x}$ would be considered irrecoverable since there are fewer equations than unknowns. However, if (i) most of the coefficients of x are zero and very few are non-zero (we say that $\mathbf{x}$ is "sparse"), and (ii) $\mathbf{A}$ satisfies certain properties, we can actually recover $\mathbf{x}$ accurately or exactly using a relatively simple convex optimization [14].

As a slight extension, similar results hold if x is not sparse but instead has a sparse frequency representation, i.e., we may write $\mathbf{x} = \mathbf{F}^{-1}\alpha$ where $\mathbf{F}^{-1}$ represents the inverse discrete Fourier transform (DFT) [5, 15, 16]. From standard intuition in classical signal processing, one would need to sample at the Nyquist rate—two times the bandwidth of the

signal [13]. Again, CS literature tells us that actually many fewer measurements are needed. Despite tremendous under-sampling in the classical sense, reliable reconstructed is possible and tractable. Numerous methods have been proposed, from linear programming strategies such as basis pursuit [14, 16, 17] and the Dantzig selector [18], first-order methods such as iterative hard thresholding [19], greedy methods such as matching pursuit [20], orthogonal matching pursuit (OMP) [21, 22] and CoSaMP [23].

We stress that this linear formulation is quite general and many real-world measurement systems can be modeled in this way [13, 24, 25]. To Illustrate this phenomenon at its most extreme, a team created a single-pixel camera which is capable of obtaining images using only a single photo-sensitive element [6]. Although this particular application is impractical because tens-of-megapixel sensors are extremely cheap, the implications of this breakthrough are great. In some applications, such as magnetic resonance imaging (MRI) [26], magnetoencephalography (MEG) [27], and hyper-spectral imaging (HSI) [28], sensing elements are very expensive whereas computation is increasingly cheap. In these situations, it is beneficial to reduce the sensing "cost," at the expense of more advanced recovery techniques.

One might imagine that CS would require a very sophisticated measurement strategy. Instead, the acquisition process is conceptually quite simple. In fact, CS tells us sensing can be done in a completely signal-agnostic and non-adaptive way by using *random projections* [29], which essentially create a compressed representation of the entire signal space. A commonly discussed sufficient condition on $\mathbf{A}$ is the *restricted isometry property* (RIP), which supposes the magnitude of sufficiently sparse vectors is approximately preserved by the given linear mapping. A matrix $\mathbf{A}$ has RIP with constant δ when

$$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta)\|\mathbf{x}\|_2^2$$

for all k-sparse vectors $\mathbf{x}$. Under this condition with $\delta_{2k} < \sqrt{2}-1$, one can show the solution $\hat{\mathbf{x}}$ to a particular convex recovery program satisfies (see e.g., [30]),

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq Ck^{-1/2}\|\mathbf{x} - \mathbf{x}_k\|_1 + C\epsilon,$$

where $\mathbf{x}_k$ is the best k-sparse approximation to x and ϵ bounds the norm of the noise $\|\mathbf{w}\|_2$.

In a way, the non-adaptive aspect of CS lies orthogonal to the aims of this thesis. Indeed, non-adaptivity is frequently discussed as one of the primary advantages of compressive sensing [13]. Even some of the earliest papers on CS discuss and dismiss the possibility of adaptive strategies [11]. Despite a decade of research, the question of adaptivity remains remains extremely nuanced, as discussed in the next section. Furthermore, despite its success, some shortcomings are known in the CS line of work:

First, the best known sensing strategies are entirely *random* (e.g., Gaussian), and so reconstruction involves working with an extremely dense matrix. Although it is known that many that many classes of randomized measurements have RIP with high probability [31], but verifying any particular sensing matrix to have the RIP is hard [32]. In many applications, random Gaussian measurements or matrices which satisfy RIP is a very unrealistic expectation [33] since we may be subject to physical measurement constraints [34]. Unfortunately, *deterministic* strategies do not approach the guarantees of random ones. Except for recent minor breakthroughs [35] there is a practical "square root bottleneck," which means on the order of k^2 deterministic measurements are required to guarantee recovery of k-sparse signals instead of about $k \log(n/k)$ as is possible in random approaches [36].

Second, compressive sensing is inherently wasteful of sensing energy. Each measurement offers low signal-to-noise ratio (SNR) [37] since the sensing effort is spread over a large number of coefficients, even if we know a-priori only a few are important. In some contexts, this is not a shortcoming, but merely a trade-off relative to other considerations of the particular application (the total number of measurements may be more important than energy, for instance). However, there may be a benefit in adaptive strategies in applications with moderate to large SNR [1].

1.2.2 Adaptivity

Soon after the application of compressive sensing was realized, researchers began considering sequential, streaming, active, and adaptive extensions (see e.g., [2, 3, 38–45]). These

approaches are most often called *adaptive compressed sensing* or *adaptive sensing*. In this context, "adaptive" means having the ability to select future measurements given the results of previous observations. This is superficially similar, but different, from the precise notion of an "adaptive estimator" in the statistics literature, which refers to an estimator which is capable of effectively estimating a parameter of interest even in the presence unknown "nuisance" parameters of which we have no interest—as well as we could if the parameters were known [46].

The central question of adaptive sensing is the following: Do measurements acquired adaptively (i.e., in consideration of previous measurements) allow one to produce a signal estimate which is "much better" than what is possible using non-adaptive type of acquisition associated with standard CS? Allowing a measurement system to intelligently choose measurements based on previously retrieved information intuitively "should" allow great benefits to sensing, but this is not true without much further qualification [37]. For instance, it is known that if a coefficient has magnitude below a particular threshold, no sensing strategy (adaptive or otherwise) can detect it with substantially less effort than standard non-adaptive methods can [1, 47].

On the other hand, several algorithms have been developed which are shown to have much better recovery than compressed sensing for sufficiently strong signals (e.g. [2, 3]). As a concrete example, the CASS algorithm from [4] succeeds in finding the support of a k-sparse signal with probability at least $1 - \delta$ provided

$$\|\mathbf{A}\|_F^2 \geq C \frac{n}{x_{\min}^2} \left(\log k + \log \left(\frac{8}{\delta} \right) \right),$$

where $x_{\min} = \min\{x_i \mid x_i \neq 0\}$. For comparison, an analogous result from non-adaptive compressed sensing requires $\|\mathbf{A}\|_F^2 \geq C' \left(n / x_{\min}^2 \right) \log n$ for accurate support recovery [4]. The adaptive procedure successfully locates the support with $O(\log(n/k))$ less sensing energy than non-adaptive schemes. Because a reliable estimate of the signal support allows the signal intensity to be estimated using classical least-square methods, left-over sensing

energy leads to a reduction in total recovery error.

1.2.3 Adaptivity in other fields

Adaptivity is seen in many fields and is by no means a new idea. In some problems, the benefit to active and sequential acquisition has been recognized for a long time (e.g., in the design of experiments [48, 49], and sequential testing [50–53]). In signal processing, it is well-established that feedback does not improve the (asymptotic) channel capacity of Gaussian channels [1, 54], although it leads to simpler encoders and decoders and makes communication faster in a non-asymptotic sense [55]. Efficient feedback in communication has recently received a renewed interest in the study of *posterior matching*, applicable to binary, Gaussian, and more general channels [56, 57].

An exceedingly general model for many problems which involve sequentially choosing actions based on noisy or partial feedback is the *partially-observable Markov decision process* (POMDP). Although solving most instances of this problem is intractable, significant research effort has to obtaining approximate solutions using dynamic programming [58]. The POMDP model has been applied to adaptive sensing in [59]. This framework also lends itself to extensions for sensing time-varying and dynamic signals [60, 61]. Unfortunately, this approach is extremely inefficient in high dimensions. Specialized approaches must be considered in the hopes of obtaining practical methods for adaptive sensing.

In many problems, it is assumed that a system making decisions does not know a-priori which actions are good or bad, but instead receives some kind of feedback, ("reward") after taking an action. At any stage the system has the option to *exploit*, taking actions seen to be good (but probably not optimal), or *explore*, searching for actions that may be better. Such an exploration/exploitation trade-off is exemplified in the *multi-armed bandit* abstraction which likens actions and rewards to pulls of a slot machine; see e.g., [62, 63].

1.2.4 1-bit and quantized compressive sensing

One aspect which was ignored in the previous section on compressive sensing, is that a real system cannot generally take and store analog, infinite-precision measurements. In digital

signal processing, front-end hardware generally uses an analog-to-digital converter. This may be modeled as introducing a quantizing function into the measurement process. We let $Q: \mathbb{R}^m \to \mathcal{Y}$ where $\mathcal{Y}$ is some discrete set:

$$\mathbf{y} = Q(\mathbf{A}\mathbf{x} + \mathbf{w} + \boldsymbol{\tau}) \in \mathcal{Y}^m$$

Here, $\mathbf{w} \in \mathbb{R}^m$ represents *pre-quantization noise* which encompasses incoming analog noise or perhaps deficiencies in analog hardware which may cause errors in digital representation. We also include $\boldsymbol{\tau} \in \mathbb{R}^m$ which is a vector of *thresholds* provided to the analog-to-digital hardware prior to quantization, which will be discussed shortly.

In some applications, we have so much granularity, say, 64 bits of quantization, that quantization is practically not an issue. In other applications, the total number of bits one is able to acquire, store, or transmit is fixed, and taking more frequent measurements, at a lower bit-depth may be preferable to fewer at a higher bit-depth [64]. In some applications, incoming signals have a tendency to saturate the analog hardware and in this case, treating the digital representation as highly quantized can lead to more effective recovery [65]. In yet other applications, quantities may be intrinsically discretized, for instance when we receive feedback in the form of star ratings or pair-wise comparisons between items in recommender system [66].

When $\mathcal{Y} = \{-1, 1\}$, each measurement may be represented by a single bit. In a sense, this is the most extreme form of quantization possible. It is initially somewhat surprising that one can recover x at all. We will often refer to this problem as *1-bit compressive sensing*. While there has been literature in higher forms of quantized compressive sensing (e.g., [67, 68]), much of the work has focused on the 1-bit case, in part to make analysis easier and as a demonstration that sensing under heavy quantization is possible. Early 1-bit CS papers, e.g., [68–71], did not use a threshold and outlined drawbacks of this model in both (i) the number of measurements required and (ii) that while the direction of x could be accurately estimated, the length of x could not be determined *at all*. More recent work has introduced

the concept of the threshold $\tau_i \neq 0$, which has made it possible to estimate $\mathbf{x}$ in both direction and norm [72, 73]. Further literature has discussed the gains available when the threshold may be chosen *adaptively* [74–76].

Restricting measurements to a single bit is actually not much of a limitation, provided the front end hardware is capable of providing a threshold. It is easy to intuit that q bits of quantization can be roughly thought of a set of q one bit quantization with varying threshold. The word "compressive" will have to be redefined a bit. We'll no longer require that the number of measurements be small compared to the signal dimension. Instead we might hope the number of measurements is close to the number of bits required to represent the signal in compressed form, which may be much smaller than the extrinsic dimension [64].

1.2.5 Learning theory

Incorporating feedback into machine learning is an extremely broad area, but the focus of this section is active learning as an extension of the probably approximately correct (PAC) framework of learning. To set the stage, we begin with fully supervised, passive (i.e., not active) machine learning. Suppose a learner would like to learn the behavior of an unknown function $f : \mathscr{A} \to \mathscr{Y}$. For simplicity, we assume $\mathscr{Y} = \{-1, 1\}$ as in "classification" problems (this fits nicely with the 1-bit CS framework[1]). To do this, the learner receives a set of m examples

$$D = (\mathbf{a}_i, y_i) \subset \mathscr{A} \times \mathscr{Y},$$

drawn from some distribution $\mathscr{D}$. The goal is to produce a $\widehat{f}$ which is somehow "close" to f. In the absence of any additional outside information, the learner will simply aim to reduce its error on the set it receives. We will further assume that the learner holds a set of possible "hypotheses" $\mathscr{F}$ and will choose the hypothesis which minimizes the number of

[1] A summary of the corresponding terms used in signal processing and machine learning: (Signal ↔ Hypothesis/concept), (Signal Structure ↔ Hypothesis/concept class), (Measurements ↔ Labeled examples, training data), (Measurement vector ↔ Unlabeled examples/instances)

incorrect outputs,

$$\widehat{f} = \arg\min_{f \in \mathscr{F}} \left| \left\{ i \;\middle|\; \widehat{f}(\mathbf{a}_i) \neq y_i \right\} \right|.$$

This is called *empirical risk minimization* [77]. The overall quality of the estimate will be measured over the entire distribution $\mathscr{D}$ (rather than the training set D);

$$\mathrm{err}(\widehat{f}) = \mathbb{P}_{\mathbf{a},y\sim\mathscr{D}}\left[\widehat{f}(\mathbf{a}) \neq y\right].$$

The ability for the learner to accomplish this (computational issues aside) will depend on a few factors, such as the number of examples available and some notion "complexity" of the set $\mathscr{F}$ (which can be measured in a variety of ways). Note that there is no assumption here of any underlying variables which may parameterize f and the space $\mathscr{F}$. The number of parameters required in a given representation is just one particular notion of complexity—another popular one is called the *Vapnik Chervonenkis* (VC) dimension. The VC dimension $\mathrm{VC}(\mathscr{F})$ measures the largest number of points which functions in $\mathscr{F}$ can assign different labels to (see e.g., [78]). The connection between 1-bit CS and classification has been noticed in e.g., [79–81].

1.2.6 Active learning

An *active* learner refers to a learner with the additional power to ask an oracle for labels to unlabeled data points $\mathbf{a}_i$. Although the learner may receive or have access to many unlabeled examples, it need not need all or even most of the labels. Clearly, if one were comparing the total number of examples, an active learner could not perform better than a *passive* one; it has access to less information since it is missing some of the labels. However, the number of *labeled* examples required may be much lower than the number of total examples, i.e., the number of labels used in the passive setting. The idea is that unlabeled data is cheap, whereas labels which require human feedback, expert knowledge, or the measurement of physical phenomena are greatly more expensive. See [82–84] for a set of excellent surveys and introductions to active learning.

A major focus of recent active learning literature has been determining in what cases

there is actually any benefit to learning actively. It is quite easy to show simple, non-pathological examples where there is no benefit to actively acquiring labels, Specifically, there are configurations of m example points where any active learner would have to query all m points [85]. In other situations, active learning has shown *exponential* reductions in the number of samples necessary to achieve the same error as passive learning. For example, when $\mathbf{a}_i$ are drawn from a log-concave distribution, [86] shows the number of labels sufficient to learn a linear separator with accuracy $\text{err}(\widehat{f}) \leq \epsilon$ and probability at least $1 - \delta$ can be reduced ($1/\epsilon$ becomes $\log 1/\epsilon$) to

$$O\Big[\Big(\log \frac{1}{\epsilon}\Big)\Big(\text{VC}(\mathscr{F}_H) + \log \frac{1}{\delta} + \log\log \frac{1}{\epsilon}\Big)\Big].$$

Broadly speaking, there are two extremes in approach seen in the literature: (i) "mellow learning" in which the learner considers examples sequentially, and for each example, queries the oracle if the corresponding label cannot be predicted with desired guaranteed accuracy and (ii) "aggressive learning," where the learner attempts to request labels which it deems "most important." The distinction of these approaches becomes clear when we consider two instances in which the aggressive approach is infeasible. One is in streaming settings where we may only observe one example at a time and once passing a label up, we cannot go back to it. Aggressive methods are also not applicable when there is no structure or prior information in the problem which we can leverage to determine which examples are important from limited feedback [87].

Most active learning methods maintain, either implicitly or explicitly, a set of possible hypotheses called the *version space*. Each piece of new information has the potential to reduces the size of the version space. operates on linear separators Although in binary classification, the hypothesis is usually thought of as a hyperplane which separates the two classes of example points, using duality we can instead treat the hypotheses as single points and each label as giving us information about which side of a particular hyperplane a hypothesis lies on. The part of the version space on the opposite side may then be discarded.

The aggressive approach of [87] aims maximize the amount of volume discarded after each label. Since labels are not known prior to asking the oracle, this amounts to maximizing the smallest side of the cut, or *bisecting* the version space as evenly as possible to further reduce the estimate uncertainty.

1.2.7 Slicing approaches

The limitation to binary feedback leads to a conceptually appealing (and occasionally optimal) approach towards localizing (i.e., estimating) a point. Since each measurement is identified with a hyperplane which divides space in two, it ought to be possible to reduce the estimation error *exponentially quickly* in the number of measurements. This concept is analogous to the "binary search" in one-dimension. It is applied to adaptive sensing as the *compressive binary search* in [88] and CASS in [4]. Another generalization of binary search to higher dimensions is discussed in [89] and extended to noisy settings in [90].

Unfortunately, this idea is complicated in a few ways when it is applied it to some problems of interest: (*i*) when pre-quantization noise is introduced, it can no longer be ensured that hyperplane cuts are accurate, (*ii*) some interesting classes of signals cannot be assigned a volume, such as the set of sparse vectors in 1-bit CS, and (*iii*) determining optimal hyperplane cuts is very difficult in high dimensions. In fact, it is difficult to test how equally a given hyperplane slices a convex set, even without noise. Since a hyperplane cuts a convex set into two disjoint convex sets, to determine how well a hyperplane bisects the version space one would merely need to compute the volumes on either side of the cut and check their ratio. However, computing the volume of a convex set is hard [91]. Luckily, there are randomized algorithms to approximately compute volumes and find good slicing hyperplanes with high probability [92].

The idea of iteratively slicing a volume to approximately localize a point is also seen in optimization where it is referred to as the *cutting plane method* [93]. In this case, it is assumed that one cannot freely choose arbitrary hyperplanes but may only interact with a *separation oracle*. The oracle knows a small set inside which one would like find a feasible

point (or determine that the problem is infeasible). When presented with a candidate point, the oracle either declares this point is inside the feasible set, or returns a hyperplane which separates the point from the feasible set. Presenting the oracle with the *centroid* of the current region would guarantee rejection of at least $1/e$ of the volume [91]. However, as with volume, the centroid is difficult to compute. Instead, one must use approximations to the centroid such as those derived from random sampling [87].

CHAPTER 2

CONSTRAINED ADAPTIVE SENSING

Suppose that we wish to estimate a vector $\mathbf{x} \in \mathbb{C}^n$ from a small number of noisy linear measurements of the form $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$, where $\mathbf{z}$ represents measurement noise. When the vector $\mathbf{x}$ is sparse, meaning that it has only s nonzeros with $s \ll n$, one can obtain a significantly more accurate estimate of $\mathbf{x}$ by adaptively selecting the rows of $\mathbf{A}$ based on the previous measurements provided that the signal-to-noise ratio (SNR) is sufficiently large. In this chapter we consider the case where we wish to realize the potential of adaptivity but where the rows of $\mathbf{A}$ are subject to physical constraints. In particular, we examine the case where the rows of $\mathbf{A}$ are constrained to belong to a finite set of allowable measurement vectors. We demonstrate both the limitations and advantages of adaptive sensing in this constrained setting. We prove that for certain measurement ensembles, the benefits offered by adaptive designs fall far short of the improvements that are possible in the unconstrained adaptive setting. On the other hand, we also provide both theoretical and empirical evidence that in some scenarios adaptivity does still result in substantial improvements even in the constrained setting. To illustrate these potential gains, we propose practical algorithms for constrained adaptive sensing by exploiting connections to the theory of optimal experimental design and show that these algorithms exhibit promising performance in some representative applications.

2.1 Introduction

Suppose that we wish to estimate a sparse vector from a small number of noisy linear measurements. In the setting where the measurements are selected in advance (independently of the signal) we now have a rich understanding of both practical algorithms and the theo-

Material in this section is joint work with Mark Davenport, Deanna Needell, and Tina Woolf. It has appeared in pre-print [94] and has lead to publications [34, 95].

retical limits on the performance of these algorithms. A typical result from this literature states that for a suitable measurement design, one can estimate a sparse vector with an accuracy that matches the minimax lower bound up to a constant factor [37]. Such results have had a tremendous impact in a variety of practical settings. In particular, they provide the mathematical foundation for "compressive sensing," a paradigm for efficient sampling that has inspired a range of new sensor designs over the last decade.

A distinguishing feature of the standard compressive sensing paradigm is that the measurements are *nonadaptive*, meaning that a fixed set of measurements are designed and acquired without allowing for any possibility of adapting as the measurements begin to reveal the structure of the signal. While this can be attractive in the sense that it enables simpler hardware design, in the context of sparse estimation this also leads to some clear drawbacks. In particular, this would mean that even once the acquired measurements show us that portions of the signal are very likely to be zero, we may still expend significant effort in "measuring" these zeros! In such a case, by *adaptively* choosing the measurements, dramatic improvements may be possible.

Inspired by this potential, recent investigations have shown that we can often acquire a sparse (or compressible) signal via far fewer measurements or far more accurately if we choose them adaptively (e.g., see [4, 45, 88, 96]). This body of work, which will be discussed in greater detail in Section 2.1.2, demonstrates that adaptive sensing indeed offers the potential for dramatic improvements over nonadaptive sensing in many settings. However, the existing approaches to adaptive sensing, which rely on being able to acquire *arbitrary* linear measurements, cannot be applied in most real-world applications where the measurements must respect certain physical *constraints*. In this chapter, our focus is on *constrained adaptive sensing*, where our measurements are restricted to be chosen from a particular set of allowable measurements. We will see that new algorithms are required and explore the theoretical limits within this more restrictive setting. Before describing the constrained adaptive setting in more detail, we first provide a brief review of existing approaches to

nonadaptive and adaptive sensing of sparse signals.

2.1.1 Nonadaptive sensing

In the standard nonadaptive compressive sensing framework [10–12, 15], we acquire a signal $\mathbf{x}$ via the linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$, where $\mathbf{A}$ is an $m \times n$ matrix representing the sensing system and $\mathbf{z}$ represents measurement noise. The goal is to design $\mathbf{A}$ so that m is smaller than n by exploiting the fact that $\mathbf{x}$ is *sparse* (or nearly sparse). Given a basis $\mathbf{\Psi}$, we say that a signal $\mathbf{x} \in \mathbb{C}^n$ is s-sparse if it can be represented by a linear combination of just s elements from $\mathbf{\Psi}$, i.e., we can write $\mathbf{x} = \mathbf{\Psi}\boldsymbol{\alpha}$, with $\|\boldsymbol{\alpha}\|_0 \leq s$, where $\|\boldsymbol{\alpha}\|_0 := |\mathrm{supp}(\boldsymbol{\alpha})|$ denotes the number of nonzeros in $\boldsymbol{\alpha}$. We will typically be interested in the case where $s \ll n$.

There is now a rich literature that describes a wide range of techniques for designing an appropriate $\mathbf{A}$ and efficient algorithms for recovering $\mathbf{x}$. In much of this literature, the matrix $\mathbf{A}$ is chosen via randomized constructions that are known to satisfy certain desirable properties such as the so-called *restricted isometry property* (RIP).[1] Under the assumption that $\mathbf{A}$ satisfies the RIP (or that $\mathbf{A}\mathbf{\Psi}$ satisfies the RIP in the case where $\mathbf{\Psi} \neq \mathbf{I}$), if each entry of $\mathbf{z}$ is independent white Gaussian noise with variance σ^2 then one can show that techniques based on ℓ_1-minimization produce an approximation $\hat{\mathbf{x}}$ satisfying

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq C \frac{n \log n}{\|\mathbf{A}\|_F^2} s\sigma^2, \tag{2.1}$$

where $C > 1$ is a fixed constant (e.g., see [12, pp. 35]). Note that this bound holds for *any* $\mathbf{x}$, and hence any SNR (even the worst-case). It is possible to obtain improved bounds that eliminate the $\log n$ factor when one assumes that the SNR is sufficiently large to ensure that the support is exactly recovered.

One can show that this result is essentially optimal in the sense that there is no alternative

[1] See Section 2.2 for a more detailed discussion of the RIP and its implications in the context of adaptive sensing. Note that the RIP is typically stated to require $\|\mathbf{A}\mathbf{x}\|_2 \approx \|\mathbf{x}\|_2$ for all s-sparse $\mathbf{x}$, which implies a fixed scaling for the matrix $\mathbf{A}$ where $\|\mathbf{A}\|_F^2 \approx n$. To ease the comparison with results that arise in contexts with alternative scalings, in the result stated in (2.1) we make no assumption on the scaling of $\mathbf{A}$ and merely require $\|\mathbf{A}\mathbf{x}\|_2 \approx \beta\|\mathbf{x}\|_2$ for some $\beta > 0$.

method to choose $\mathbf{A}$ or perform the reconstruction that can do better than this (up to the precise value of the constant C) [37]. In the event that the signal $\mathbf{x}$ is not exactly s-sparse, it is also possible to extend these results by introducing an additional term in the error bound that measures the error incurred by approximating $\mathbf{x}$ as s-sparse. See [12] and references therein for further details.

2.1.2 Adaptive sensing

A defining feature of the approach described above is that it is completely nonadaptive. When we consider the effect of noise, this nonadaptive approach might draw some severe skepticism. To see why, note that in the nonadaptive scenario, most of the "sensing energy" is used to measure the signal at locations where there is no information, i.e., where the signal vanishes. Specifically, one consequence of using the randomized constructions for $\mathbf{A}$ typically considered in the literature, or alternatively, any matrix satisfying the RIP, is that the available sensing energy (i.e., $\|\mathbf{A}\|_F^2$) is evenly distributed across all possible indices. This is natural since *a priori* we do not know where the nonzeros may lie, however, since most of the coordinates $\mathbf{x}_j$ are zero, it also means that the vast majority of the sensing energy is seemingly wasted. In other words, by design, the sensing vectors are approximately orthogonal to the signal, yielding a poor signal-to-noise ratio (SNR).

The idea behind adaptive sensing is that we should focus our sensing energy on locations where the signal is nonzero in order to increase the SNR, or equivalently, not waste sensing energy. In other words, one should try to learn as much as possible about the signal while acquiring it in order to design more effective subsequent measurements. Roughly speaking, one would like to (i) detect those entries which are nonzero or significant, (ii) progressively concentrate the sensing vectors on those entries, and (iii) estimate the signal from such localized linear functionals. Such a strategy is employed by the *compressive binary search* and *compressive adaptive sense and search* strategies of [88] and [4]. These algorithms operate by examining successively smaller pieces of the signal to accurately determine the locations of signal energy. These techniques can yield dramatic improvements in recovery

accuracy.

To quantify the potential benefits of an adaptive scheme, suppose that we observe

$$y_i = \langle \mathbf{a}_i, \mathbf{x} \rangle + z_i \tag{2.2}$$

where the z_i are independent and identically distributed (i.i.d.) $\mathcal{N}(0, \sigma^2)$ entries and the $\mathbf{a}_i$ are allowed to depend on the measurement history $((y_1, \mathbf{a}_1), \cdots, (y_{i-1}, \mathbf{a}_{i-1}))$, with the only constraint being that $\sum_i \|\mathbf{a}_i\|_2^2 = \|\mathbf{A}\|_F^2$ is fixed. Consider a simple procedure that uses half of the sensing energy in a nonadaptive way to identify the support of an s-sparse vector $\mathbf{x}$ and then adapts to use the remaining half of the sensing energy to estimate the values of the nonzeros. If such a scheme identifies the correct support, then it is easy to show that this procedure can yield an estimate satisfying

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 = \frac{2s}{\|\mathbf{A}\|_F^2} s\sigma^2. \tag{2.3}$$

If we contrast this result to that in (2.1), which represents the best possible performance in the nonadaptive setting, we see that this simple adaptive scheme can potentially improve upon the nonadaptive scheme by a factor of roughly $(n/s) \log n$, which represents a *dramatic* improvement in the typical scenario where $s \ll n$. Of course, this is predicated on the assumption that the first stage of support identification succeeds, which is not always the case.

A fundamental question is thus: *in practice, how much lower can the mean squared error (MSE) be when we are allowed to sense the signal adaptively*? The answer is a subtle one. In [1], it is shown that there is a fixed constant $C > 0$ such that

$$\inf_{\hat{\mathbf{x}}} \sup_{\|\mathbf{x}\|_0 \le s} \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \ge C \frac{n}{\|\mathbf{A}\|_F^2} s\sigma^2. \tag{2.4}$$

In other words, for even the best possible adaptive scheme there are s-sparse vectors for which our recovery error is bounded below by (2.4). This lower bound improves upon the nonadaptive performance (2.1) by only a factor of $\log n$, coming far short of the improvement

that (2.3) indicates might be possible. Similar results are also obtained in [47]. These results are established by considering vectors that are so difficult to estimate that it is impossible to obtain a reliable estimate of their support, and so adaptive algorithms offer limited room for improvement over nonadaptive ones.

The result (2.4) does not say that adaptive sensing *never* helps. In fact, in practice it *almost always* does help. For example, when some or most of the nonzero entries in $\mathbf{x}$ are only slightly larger than the worst-case amplitude identified in [1], we *can* detect them sufficiently reliably to enable the dramatic improvements predicted in (2.3). More concretely, provided that σ^2 is not too large[2] relative to the nonzero entries of $\mathbf{x}$, a well-designed adaptive scheme, where the $\mathbf{a}_i$ are chosen sequentially as in [4, 88], can achieve

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq C' \frac{s}{\|\mathbf{A}\|_F^2} s\sigma^2 \tag{2.5}$$

for a fixed constant C', which represents an enormous improvement when $s \ll n$, and demonstrates that the potential benefits suggested in (2.3) can be realized in certain regimes.

We briefly note that these results are somewhat reminiscent of classical results from the field of information based complexity [9, 11, 97, 98] as well as more recent results in active learning [99]. Although this literature considers different observation models (e.g., noise-free observations of non-sparse signals), the general theme is that adaptivity is beneficial only in certain regimes (e.g., see [45]). In another direction, we also note that several authors have previously suggested Bayesian approaches to adaptive sensing that are highly relevant to the problems we study in this chapter, but which currently lack much in the way of theoretical justification or understanding [2, 100, 101].

2.1.3 Constrained sensing

Up to this point, we have discussed results in which we essentially have complete freedom to design both the adaptive and nonadaptive measurements in an optimal fashion (that is,

[2]For example, the compressive binary search procedure proposed in [88] succeeds in finding the location of the smallest nonzero entry of amplitude μ with probability $1-\delta$ when $\mu^2/\sigma^2 > 16n\log(\frac{1}{2\delta}+1)/\|A\|_F^2$. The result for the procedure in [4] is similar.

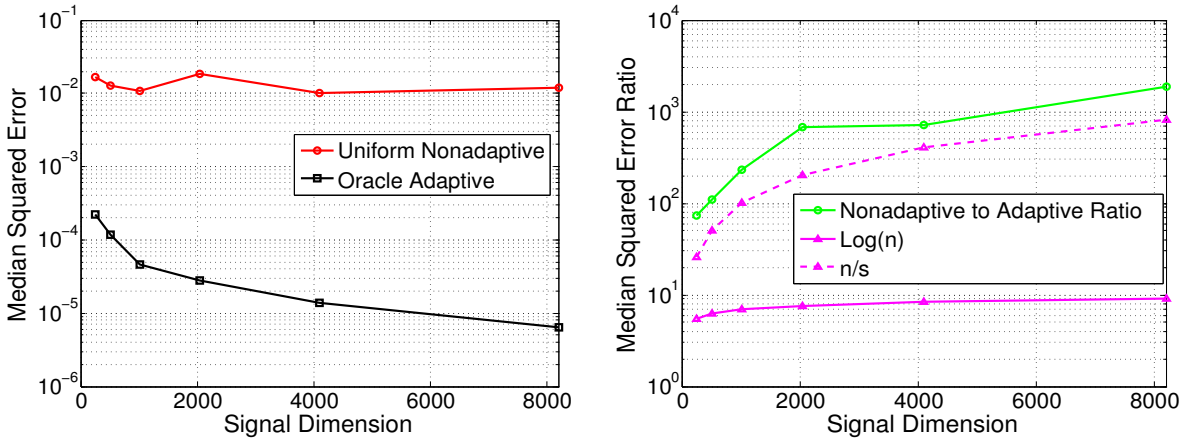


Figure 2.1: (Left) The median squared error versus the signal dimension n for nonadaptive recovery with uniformly random selected measurements (red) and oracle adaptive recovery (black). (Right) The ratio (green) of the nonadaptive median squared recovery error to the oracle adaptive median squared recovery error versus the signal dimension n, with $\log n$ (solid magenta) and n/s (dashed magenta) included for reference.

up to a constraint on $\|\mathbf{A}\|_F$). However, there are many applications where such freedom does not exist, and there are significant constraints on the kind of measurements that we can actually acquire. Such constraints arise in various hardware devices inspired by compressive sensing. For example, the single-pixel camera [102] acquires samples of an image by computing inner products with binary patterns. In this application we could still utilize adaptive measurements, but they must be binary. In other applications, we may be restricted to obtaining point samples of the signal of interest. For example, in standard sampling systems we are restricted to individually measuring each signal coefficient over time or space. Finally, in tomography and magnetic resonance imaging (MRI), as well as other medical imaging settings, we cannot acquire inner products with arbitrary linear functionals—we are limited to Fourier measurements.

In all of these settings, the measurements are *constrained*; we still have the flexibility to design measurements adaptively, but we can only select measurements from a fixed ensemble of predetermined measurements. Thus, the constrained setting will typically preclude the use of any of the adaptive sensing algorithms referenced above, and a new approach is required. Specifically, if we let $\mathscr{M} \subset \mathbb{C}^n$ denote the set of candidate measurement vectors,

then the constrained adaptive sensing problem becomes one of sequentially selecting the rows $\mathbf{a}_i$ of our sensing matrix from the set $\mathscr{M}$. In this work, we assume the multiplicity of a particular measurement from $\mathscr{M}$ is allowed to be greater than one; that is, repeated measurements are permitted. For the methods discussed in this chapter, we will restrict our attention to the case where $\mathscr{M}$ is a finite set. For a majority of our discussion and examples, we will focus on the setting where $\mathscr{M} = \{f_1, f_2, \ldots, f_n\}$ consists of rows from the Discrete Fourier Transform (DFT) matrix. We stress, however, that we need not require $|\mathscr{M}| = n$ in general.

With the restriction that the measurements be chosen from the DFT ensemble, Figure 2.1 illustrates the large potential difference between a completely nonadaptive sensing scheme, where the measurements are selected uniformly at random, and an "oracle" adaptive sensing scheme which uses a priori knowledge of the true locations of the nonzeros in a signal to carefully adapt the choice of measurement vectors to minimize the expected recovery error using the strategy outlined in Section 2.3. In both cases, the Compressive Sampling Matching Pursuit (CoSaMP) [23] algorithm is used for the signal recovery. The median[3] squared error over 200 trials is displayed against the signal dimension n. Here, the signal is chosen to have a sparse Haar wavelet decomposition that is supported on a tree. The choice of a tree-sparse signal is motivated by the observation that natural images typically have a structured sparsity pattern in a wavelet domain due to correlations between scales. In these simulations, the noise level $\sigma^2 = 10^{-4}$ is held constant while the nonzero coefficients scale as $\sqrt{n}$ so that the per-measurement SNR is fixed. The number of measurements taken is set to be $m = 0.6n$ (rounding when necessary). See Section 2.4.1 for further details regarding these simulations.

It is well-known that the DFT and Haar wavelet transforms are not incoherent, which implies that $\mathbf{A}\Psi$ should not satisfy the RIP; hence, we would not expect blind nonadaptive

[3] Note that the *median* and *mean* curves exhibit the same overall behavior; however, we display the median error across all trials rather than the mean error throughout because the median, being a more robust measure, resulted in smoother curves with clearer trends between the methods.

sensing to do well in this setting. However, Figure 2.1 does illustrate the large potential for improvement over nonadaptive sensing. In this case, the adaptive algorithm can potentially improve the recovery error over nonadaptive sensing by roughly a factor of n/s, which represents a substantial gain when $s \ll n$. While we will see below that there are also nonadaptive strategies to address the coherence of Fourier and Haar which somewhat reduce the gap between adaptive and nonadaptive sensing in this case, we believe that this clearly illustrates the potential for adaptive sensing, even in the constrained setting.

2.1.4 Organization

The remainder of the chapter is organized as follows. In Section 2.2, we show a simple lower bound on the adaptive performance of systems limited to DFT measurements. We then generalize this result to the larger class of measurements satisfying the RIP. In both cases, the signal is assumed to be sparse in the canonical basis. In Section 2.3, we give a method for measurement selection based on optimal experimental design. In Section 2.4, we provide simulations in a more realistic setting and display numerical results when Fourier measurements are used and the signal is assumed to be sparse in the Haar wavelet basis, for both synthetic and realistic signals. We also present some analytical justification using 1-sparse signals in this constrained adaptive setting. Finally, we conclude in Section 2.5 with a brief discussion.

2.2 Lower bounds on adaptive performance

The main result of this section shows that adaptive sensing cannot offer substantial improvements over the nonadaptive scheme when the measurements are restricted to certain specific classes of ensembles and the signal is sparse in the canonical basis (i.e., $\mathbf{\Psi} = \mathbf{I}$). We first consider the Fourier ensemble, where the sensing vectors are chosen from the rows of the DFT matrix $\mathbf{F} \in \mathbb{C}^{n \times n}$, where $\mathbf{F}$ has entries

$$f_{jk} = \frac{1}{\sqrt{n}} \exp(-2\pi\sqrt{-1}jk/n) \tag{2.6}$$

for $j, k = 0, 1, \ldots, n-1$. In this constrained setting we have the following lower bound.

Theorem 2.2.1. *Under the adaptive measurement model of* (2.2)*, where the* $\mathbf{a}_i$ *are chosen (potentially adaptively and allowing repeated measurements) by selecting rows from the DFT matrix* (2.6)*, we have that*

$$\inf_{\hat{\mathbf{x}}} \sup_{\substack{\|\mathbf{x}\|_0 \le s \\ \|\mathbf{x}\|_2 \ge R}} \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \ge \frac{n}{m} s\sigma^2 \tag{2.7}$$

for any $R \ge 0$*.*

This shows that even using an optimal choice of sensing vectors, the recovery error is still proportional to $\frac{n}{m}s\sigma^2$, *even if we exclude the low-SNR setting* (by setting R to be large relative to σ). This is somewhat reminiscent of the main results of [1] and [47], which (in an unconstrained setting) establish minimax bounds of the form given in (2.4). However, a key difference is that in the unconstrained setting the worst-case error which defines the minimax rate is determined by the performance at a certain range of worst-case SNRs. Specifically, these bounds are obtained by constructing a "least favorable prior" where the nonzeros of $\mathbf{x}$ are near a specific level,[4] and thus if we were to exclude these challenging $\mathbf{x}$ via the restriction that $\|x\|_2 \ge R$ as in (2.7), the bound in (2.4) would be dramatically lower – in particular, the gains shown in (2.5) could be realized [4, 88]. Thus, in a sense Theorem 2.2.1 is far more pessimistic than these results since it applies *no matter how large the SNR* – although given the incoherence of the DFT and the canonical bases, perhaps this is not that surprising. Finally, we note that for certain values of R it may be possible to obtain a slightly stronger version of Theorem 2.2.1 (by a $\log n/\log\log n$ factor) using the techniques in [103, Thm 6.1]. We do not pursue these refinements here.

Proof of Theorem 2.2.1. For any adaptive procedure $\hat{\mathbf{x}}$, we let $\mathbf{F}'$ be the $m\times n$ sensing matrix consisting of the m adaptively chosen vectors from the rows of $\mathbf{F}$, and let $\mathbf{F}'_\Lambda$ denote the $m\times s$ submatrix of $\mathbf{F}'$ whose column indices correspond to the indices of the support Λ of $\mathbf{x}$. Using the rows of $\mathbf{F}'$ to acquire the measurements as in (2.2), we obtain $\mathbf{y} = \mathbf{F}'\mathbf{x} + \mathbf{z} = \mathbf{F}'_\Lambda \mathbf{x}_\Lambda + \mathbf{z}$.

[4] This threshold is around $(\min_i x_i^2)/\sigma^2 \approx (n/m)\log s$.

It is not difficult to show (e.g., see the Appendix of [37]) that

$$\inf_{\hat{\mathbf{x}}} \sup_{\substack{\|\mathbf{x}\|_0 \leq s \\ \|\mathbf{x}\|_2 \geq R}} \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \geq \inf_{\hat{\mathbf{x}}} \sup_{\substack{\mathbf{x}' \in \mathbb{R}^s \\ \|\mathbf{x}'\|_2 \geq R}} \mathbb{E}\|\hat{\mathbf{x}}(\mathbf{F}_\Lambda' \mathbf{x}' + \mathbf{z}) - \mathbf{x}'\|_2^2,$$

where $\hat{\mathbf{x}}(\cdot)$ takes values in $\mathbb{R}^s$.

To establish the bound in (2.7) we consider a sequence of least favorable prior distributions on $\mathbf{x}'$. The minimax risk is always larger than the Bayes risk under any prior, so this will establish a lower bound on the minimax risk. Towards this end, consider the prior on $\mathbf{x}'$ where $\mathbf{x}' \sim \mathcal{N}(0, \rho^2 \mathbf{I})$, but where the distribution is truncated to be zero for $\|\mathbf{x}'\|_2 \leq R$ and re-scaled appropriately. Note that in the absence of this truncation, the Bayes risk would be given by

$$\sigma^2 \sum_{i=1}^{s} \left(\frac{\sigma_i(\mathbf{F}_\Lambda')}{\sigma_i^2(\mathbf{F}_\Lambda') + \frac{\sigma^2}{\rho^2}} \right)^2, \tag{2.8}$$

where $\sigma_i(\mathbf{F}_\Lambda')$ denotes the i^{th} singular value of $\mathbf{F}_\Lambda'$. This follows from the fact that the Bayes estimator is given by

$$\mathbb{E}[\mathbf{x}|\mathbf{y}] = (\mathbf{F}_\Lambda'^T \mathbf{F}_\Lambda' + \tfrac{\sigma^2}{\rho^2}\mathbf{I})^{-1} \mathbf{F}_\Lambda'^T \mathbf{y}.$$

The result in (2.8) follows from the fact that for this estimator the expected squared error is given by

$$\sigma^2 \|(\mathbf{F}_\Lambda'^T \mathbf{F}_\Lambda' + \tfrac{\sigma^2}{\rho^2}\mathbf{I})^{-1} \mathbf{F}_\Lambda'^T\|_F^2$$

which reduces to (2.8) via the application of standard properties of the singular value decomposition. We now note that for any $R \geq 0$, as $\rho^2 \to \infty$, the Bayes risk for the truncated prior will converge to that of (2.8), namely,

$$\sigma^2 \|(\mathbf{F}_\Lambda'^T \mathbf{F}_\Lambda')^{-1} \mathbf{F}_\Lambda'^T\|_F^2 = \sigma^2 \sum_{i=1}^{s} \frac{1}{\sigma_i^2(\mathbf{F}_\Lambda')}$$

Putting this all together, we have that

$$\begin{aligned}\inf_{\hat{\mathbf{x}}} \sup_{\substack{\|\mathbf{x}\|_0 \le s \\ \|\mathbf{x}\|_2 \ge R}} \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 &\ge \sigma^2 \sum_{i=1}^{s} \frac{1}{\sigma_i^2(\mathbf{F}'_\Lambda)} \\ &\ge \sigma^2 \frac{s^2}{\sum_{i=1}^{s} \sigma_i^2(\mathbf{F}'_\Lambda)} \\ &= \sigma^2 \frac{s^2}{\|\mathbf{F}'_\Lambda\|_F^2}\end{aligned}$$

where the second inequality follows from Jensen's inequality. Since $\|\mathbf{F}'_\Lambda\|_F^2 = \frac{sm}{n}$, this completes the proof. □

Our next result generalizes this type of lower bound to any ensemble whose submatrices satisfy the RIP with overwhelming probability. This statement is significant because it suggests that in some constrained situations, specifically many commonly studied in compressive sensing, there is little benefit from adaptivity. Formally, we define an RIP ensemble as follows.

Definition 2.2.2. *Let m be fixed. We say that an $n \times n$ matrix $\mathbf{A}$ with unit-norm rows is an* RIP ensemble *if for any $m' \ge m$ a random $m' \times n$ submatrix $\widetilde{\mathbf{A}}$, whose rows are uniformly chosen without replacement, satisfies*

$$0.5\frac{m'}{n}\|\mathbf{u}\|_2^2 \le \|\widetilde{\mathbf{A}}\mathbf{u}\|_2^2 \le 1.5\frac{m'}{n}\|\mathbf{u}\|_2^2, \tag{2.9}$$

for all s-sparse $\mathbf{u}$ with probability $1 - \exp(-cn)$ (where c is such that $\exp(-cn) < 1/2n$).

Theorem 2.2.3 makes rigorous the claim that selecting rows intelligently from such a matrix yields no substantial improvement over a nonadaptive scheme.

Theorem 2.2.3. *Under the adaptive measurement model of* (2.2)*, where the $\mathbf{a}_i$ are chosen (potentially adaptively and allowing repeated measurements) by selecting rows from an RIP*

ensemble as defined above, we have that

$$\inf_{\hat{\mathbf{x}}} \sup_{\substack{\|\mathbf{x}\|_0 \le s \\ \|\mathbf{x}\|_2 \ge R}} \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \ge \frac{sn}{3m^2} s\sigma^2 \tag{2.10}$$

for any $R \ge 0$.

We note that one usually anticipates m to be on the order of $s \log n$, in which case this bound becomes

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \ge \frac{sn}{3ms \log n} s\sigma^2 = \frac{n}{3m \log n} s\sigma^2,$$

which is roughly a factor of $\log^2 n$ lower than the upper bound in (2.1). This result shows that the recovery error with any adaptive measurements selected from some standard RIP ensemble again falls short of the possible gains shown in (2.3).

We also note here that the bound in Theorem 2.2.3 is worse by a factor of m/s than Theorem 2.2.1. However, we believe this is necessary due to the fact that the only assumption we place on $\mathbf{A}$ is that (2.9) holds with overwhelming probability; this is a much weaker requirement than insisting on DFT measurements as in Theorem 2.2.1. As a motivating example, fix some subset $\Lambda \subset \{1, \dots, n\}$ of size s. Construct a matrix $\mathbf{A}$ by setting it to the DFT basis $\mathbf{F}$, with its first row modified in the following way: on Λ, multiply each entry by a factor of C where $C^2 = m/8s$ and off of Λ multiply each entry by a factor $c = \sqrt{(n - sC^2)/(n - s)}$. This yields a matrix $\mathbf{A}$ whose rows still have unit norm. In addition, one can show that for this new matrix $\mathbf{A}$, the property (2.9) still holds with the same probability for $\delta = 5/8$ for any $(m+1) \times n$ submatrix $\widetilde{\mathbf{A}}$. Construct an $(m+1) \times n$ matrix $\mathbf{A}'$ with the first row of $\mathbf{A}$ repeated $m - s + 2$ times (since at least s rows need to be unique). Then one computes that $\|\mathbf{A}'_\Lambda\|_F^2 = \frac{s}{n}(s - 1 + \frac{m}{8s}(m - s + 2)) \gtrsim m^2/n$. On the other hand, any matrix of the same size adaptively constructed from the DFT basis $\mathbf{F}$ has a squared Frobenius norm equal to $s(m+1)/n$. Thus we may indeed lose an m/s factor because of this weakened assumption.

Proof of Theorem 2.2.3. Let $\mathbf{A}'$ be the $m \times n$ matrix of the adaptively selected rows as in the theorem. Fix a support set Λ of size at most s. Let $\mathbf{A}'_\Lambda$ be the restriction of $\mathbf{A}'$ to

the support set Λ. We will prove the result by showing a bound on the norm of the rows of $\mathbf{A}'_\Lambda$ which we obtain via an argument of contradiction. To that end, let $\mathbf{a}^\star$ be the row of $\mathbf{A}$ corresponding to the row of $\mathbf{A}'_\Lambda$ with the greatest Euclidean norm. Now consider drawing a random $(m+1) \times n$ submatrix $\widetilde{\mathbf{A}}$ of $\mathbf{A}$ that contains $\mathbf{a}^\star$ as a row. Then one can compute that any such submatrix $\widetilde{\mathbf{A}}$ satisfies (2.9) (with $m' = m+1$) with probability at least $1 - \exp(-cn)n/(m+1) > 1 - \exp(-cn)n$. Indeed, one sees formally that

$$\begin{aligned}
&\mathbb{P}(\widetilde{\mathbf{A}} \text{ does not satisfy (2.9)} \mid \mathbf{a}^\star \text{ is a row of } \widetilde{\mathbf{A}}) \\
&= \frac{\mathbb{P}(\widetilde{\mathbf{A}} \text{ does not satisfy (2.9) and } \mathbf{a}^\star \text{ is a row of } \widetilde{\mathbf{A}})}{\mathbb{P}(\mathbf{a}^\star \text{ is a row of } \widetilde{\mathbf{A}})} \\
&\leq \frac{\mathbb{P}(\widetilde{\mathbf{A}} \text{ does not satisfy (2.9))}}{\mathbb{P}(\mathbf{a}^\star \text{ is a row of } \widetilde{\mathbf{A}})} \\
&\leq \frac{\exp(-cn)}{(m+1)/n}.
\end{aligned}$$

Now let $\widetilde{\mathbf{A}}^c$ be the remainder of the matrix, i.e., all rows of $\widetilde{\mathbf{A}}$ except row $\mathbf{a}^\star$. Similarly, one computes that any such matrix $\widetilde{\mathbf{A}}^c$ satisfies (2.9) (with $m' = m$) with probability at least $1 - \exp(-cn)n/(n-m) > 1 - \exp(-cn)n$. Thus *both* of these matrices satisfy (2.9) with probability at least $1 - 2\exp(-cn)n > 0$. For the sake of a contradiction, suppose that $\|\mathbf{a}^\star_\Lambda\|_2^2 > 3m/n$. Observe that the signal $\mathbf{x} \in \mathbb{R}^n$ where $\mathbf{x}_\Lambda = \mathbf{a}^\star_\Lambda$ and padded with zeros off of the support Λ is an s-sparse signal. Then since both matrices satisfy (2.9), we must have that

$$\begin{aligned}
\|\widetilde{\mathbf{A}}\mathbf{x}\|_2^2 &= \|\widetilde{\mathbf{A}}^c\mathbf{x}\|_2^2 + |\langle \mathbf{a}^\star_\Lambda, \mathbf{x}\rangle|_2^2 \\
&\geq 0.5\frac{m}{n}\|\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2^4 \\
&> \left(0.5\frac{m}{n} + \frac{3m}{n}\right)\|\mathbf{x}\|_2^2 \geq \frac{3.5m}{n}\|\mathbf{x}\|_2^2.
\end{aligned}$$

On the other hand, we must also have that

$$\|\widetilde{\mathbf{A}}\mathbf{x}\|_2^2 \leq 1.5\frac{m+1}{n}\|\mathbf{x}\|_2^2.$$

Combining these means that $\frac{3.5m}{n} \leq 1.5\frac{m+1}{n} \leq \frac{3m}{n}$, which is a contradiction. Thus, it must be that $\|\mathbf{a}_\Lambda^\star\|_2^2 \leq 3m/n$. Since $\mathbf{a}_\Lambda^\star$ is the largest row of $\mathbf{A}'_\Lambda$, we then have that

$$\|\mathbf{A}'_\Lambda\|_F^2 \leq m\|\mathbf{a}_\Lambda^\star\|_2^2 \leq \frac{3m^2}{n}.$$

Following the same argument as in the proof of Theorem 2.2.1, we thus have that

$$\inf_{\hat{\mathbf{x}}} \sup_{\substack{\|\mathbf{x}\|_0 \leq s \\ \|\mathbf{x}\|_2 \geq R}} \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \geq \frac{s^2}{\|\mathbf{A}'_\Lambda\|_F^2}\sigma^2 \geq \frac{sn}{3m^2}s\sigma^2,$$

which completes the proof.

□

2.3 Adaptivity through optimal experimental design

Although there are some settings where constrained adaptive sensing does not offer substantial improvement over the nonadaptive scheme, one can of course ask if there are other settings where notable gains are still possible. In order to address this question, we consider the simplified constrained adaptive sensing problem where we assume the support Λ of the signal $\mathbf{x}$ (with respect to the sparsity basis $\mathbf{\Psi}$) is known, or some estimate of the support is provided. How would we choose the measurements to best make use of this information, while still respecting that the measurements are constrained to be from the measurement ensemble $\mathscr{M}$?

Let $\{\mathbf{a}_i\}_{i=1}^m$ denote a sequence of length m with elements $\mathbf{a}_i \in \mathscr{M}$ corresponding to the measurements of $\mathscr{M}$ that are chosen.[5] Then, denote by $\mathbf{A}'$ the $m \times n$ matrix (recall $\mathscr{M} \subset \mathbb{C}^n$) whose i^{th} row is $\mathbf{a}_i$. If $\Lambda = \text{supp}(x)$, then it can be shown by following the arguments in the

[5] We use a *sequence* of elements from $\{1, \ldots, |\mathscr{M}|\}$ rather than a *subset* to emphasize that the m measurements from $\mathscr{M}$ need not be distinct. Note that in the general adaptive setting the *order* of the measurements is also important; however, in the context of this section there is only one batch of adaptive measurements and thus the order within this batch has no impact.

proof of Theorem 2.2.1 that the optimal MSE satisfies

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2 &= \|(\mathbf{A}'\boldsymbol{\Psi}_\Lambda)^\dagger\|_F^2\sigma^2 \\ &= \mathrm{Tr}\left(((\mathbf{A}'\boldsymbol{\Psi}_\Lambda)^*\mathbf{A}'\boldsymbol{\Psi}_\Lambda)^{-1}\right)\sigma^2,\end{aligned} \tag{2.11}$$

where $(\mathbf{A}'\boldsymbol{\Psi}_\Lambda)^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{A}'\boldsymbol{\Psi}_\Lambda$, σ^2 is the variance of the noise term as in (2.2), $\boldsymbol{\Psi}_\Lambda$ is the submatrix of $\boldsymbol{\Psi}$ restricted to the columns indexed by Λ, and $\mathbf{A}'\boldsymbol{\Psi}_\Lambda$ is assumed to have full (column) rank. Our goal is to find a length-m measurement sequence $\{\mathbf{a}_i\}_{i=1}^m$ that minimizes (2.11), which is equivalent to solving

$$\{\hat{\mathbf{a}}_i\}_{i=1}^m = \underset{\{\{\mathbf{a}_i\}_{i=1}^m \,|\, \mathbf{a}_i\in\mathscr{M}\}}{\arg\min} \mathrm{Tr}\left(((\mathbf{A}'\boldsymbol{\Psi}_\Lambda)^*\mathbf{A}'\boldsymbol{\Psi}_\Lambda)^{-1}\right), \tag{2.12}$$

where $\mathbf{A}' = \mathbf{A}'(\{\mathbf{a}_i\}_{i=1}^m)$ is constructed as described above. Note that an essentially equivalent way to state (2.12) (up to a permutation of the measurements) is via the discrete optimization problem

$$\begin{aligned}\hat{\mathbf{S}} = \underset{\substack{\text{diagonal matrices } \mathbf{S}\succeq 0 \\ s_{ii}\in\mathbb{Z}^+}}{\arg\min} \quad & \mathrm{Tr}\left(((\mathbf{A}\boldsymbol{\Psi}_\Lambda)^*\mathbf{S}\mathbf{A}\boldsymbol{\Psi}_\Lambda)^{-1}\right) \\ \text{subject to } & \mathrm{Tr}(\mathbf{S}) \le m,\end{aligned} \tag{2.13}$$

where $\mathbf{A}$ is the $|\mathscr{M}| \times n$ matrix containing all possible measurement vectors from $\mathscr{M}$ and $s_{ii} \in \mathbb{Z}^+$ forces each diagonal entry of $\mathbf{S}$ to be a non-negative integer (reflecting the multiplicity of each $\mathbf{a}_i$). Both (2.12) and (2.13) reflect the optimization problem that we would ideally like to solve. Unfortunately they are computationally demanding discrete optimization problems; hence, we instead consider the relaxation of (2.13)

$$\begin{aligned}\hat{\mathbf{S}} = \underset{\text{diagonal matrices } \mathbf{S}\succeq 0}{\arg\min} \quad & \mathrm{Tr}\left(((\mathbf{A}\boldsymbol{\Psi}_\Lambda)^*\mathbf{S}\mathbf{A}\boldsymbol{\Psi}_\Lambda)^{-1}\right) \\ \text{subject to } & \mathrm{Tr}(\mathbf{S}) \le m,\end{aligned} \tag{2.14}$$

where the constraint $\mathrm{Tr}(\mathbf{S}) \le m$ ensures that the resulting "weighted" sensing matrix $\sqrt{\mathbf{S}}\mathbf{A}$ satisfies the "sensing energy" constraint $\|\sqrt{\mathbf{S}}\mathbf{A}\|_F^2 \le m$ when the rows of $\mathbf{A}$ are normalized.

Note that this is equivalent to the continuous design for the A-optimality criterion studied in the optimal experimental design literature [104].

Fortunately, (2.14) is a convex problem [105] and can be efficiently solved by a number of methods. Whereas the problem in (2.12) would tell us which measurements and how many of each to use from $\mathscr{M}$, (2.14) instead tells us, through the diagonal matrix $\hat{\mathbf{S}}$ of weights, "how much" of each measurement to use. We simply weight each $\mathbf{a}_i$ by $\sqrt{\hat{s}_{ii}}$, where $\hat{s}_{ii}$ denotes the i^{th} element on the diagonal of $\hat{\mathbf{S}}$.

If, on the other hand, we use the measurement model where we must choose m unweighted measurements from $\mathscr{M}$, the practical use of $\hat{\mathbf{S}}$ from (2.14) is less obvious. We experimented with several different (though likely sub-optimal) approaches to using the weights in $\hat{\mathbf{S}}$, and the following method empirically seemed to produce the best results. In this work, we use a simple sampling scheme to obtain a discrete design. Specifically, we draw exactly m measurements, with replacement, according to the probability mass function

$$p_i = \frac{\hat{s}_{ii}}{m}. \tag{2.15}$$

We guarantee that the resulting matrix $\mathbf{A}'$ is at least rank s by rejecting any construction for which this constraint is not satisfied. These m measurements then form the rows of the sensing matrix $\mathbf{A}'$.

2.4 Case study: Fourier measurements of Wavelet sparse signals

The results of Section 2.2 demonstrate that adaptive sensing cannot offer substantial improvements over nonadaptive sensing for certain classes of measurement ensembles when the signal is sparse in the canonical basis. We next explore the case when $\mathbf{\Psi}$ is instead a wavelet basis and we acquire DFT measurements (this is indeed the setting of Figure 2.1, which suggests dramatic potential improvements from constrained adaptive sensing). This setting serves as a somewhat idealized model for a number of applications in tomography and other medical imaging since physical limitations would entail that we can only acquire DFT measurements, and realistic images are generally sparse with respect to wavelet

bases [106]. In this setting we might receive one DFT measurement at a time, and from those, we can (potentially in real time) request the next DFT coefficient to be measured.

For our first two sets of experiments, we will assume the sparsity basis Ψ is the Haar wavelet basis. We will denote the $n \times n$ discrete Haar wavelet transform by $\mathbf{H}$, with entries h_{jk} for $j, k = 0, 1, \ldots, n-1$ and n is assumed to be some power of 2. When $j = 0$, we have

$$h_{0k} = \frac{1}{\sqrt{n}}. \tag{2.16}$$

For indices $j > 0$, we write $j = 2^p + q - 1$, where $p = \lfloor \log_2 j \rfloor$ and q are nonnegative integers, and define

$$h_{jk} = \frac{1}{\sqrt{n}} \begin{cases} 2^{p/2} & \frac{(q-1)n}{2^p} \leq k < \frac{(q-0.5)n}{2^p} \\ -2^{p/2} & \frac{(q-0.5)n}{2^p} \leq k < \frac{qn}{2^p} \\ 0 & \text{otherwise.} \end{cases} \tag{2.17}$$

Since, however, the Haar wavelet basis $\mathbf{H}$ is a sparsifying transformation, for a signal (or image) $\mathbf{x}$ we have that $\mathbf{Hx} = \boldsymbol{\alpha}$, with $\|\boldsymbol{\alpha}\|_0 \leq s$. This means $\mathbf{x} = \mathbf{H}^*\boldsymbol{\alpha}$, where $\mathbf{H}^*$ denotes the adjoint of $\mathbf{H}$, for which $\mathbf{H}^* = \mathbf{H}^{-1}$ since $\mathbf{H}$ is unitary.

With this notation in hand and recalling that $\mathbf{F}$ is the $n \times n$ DFT, (2.11) becomes

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 = \|(\mathbf{F}'\mathbf{H}_\Lambda^*)^\dagger\|_F^2 \sigma^2, \tag{2.18}$$

where $\mathbf{F}'$ is the $m \times n$ sensing matrix consisting of the m adaptively chosen vectors from $\mathbf{F}$ and $\mathbf{H}_\Lambda^*$ is the $n \times s$ submatrix of $\mathbf{H}^*$ restricted to the columns indexed by $\Lambda = \text{supp}(\boldsymbol{\alpha}) = \text{supp}(\mathbf{Hx})$. Thus, we see that the optimal MSE depends on the correlations of the DFT and Haar basis elements. In a similar manner, in our last experiment, where the signal is an MRI image, we will assume the sparsity basis $\boldsymbol{\Psi}$ is the Daubechies wavelet with 3 vanishing moments (D6).

We now present a suite of numerical simulations in these settings that employ the relaxation (2.14) followed by the sampling scheme described in Section 2.3 to select a sequence of m DFT measurement vectors. We then follow with a short analysis for the simple case

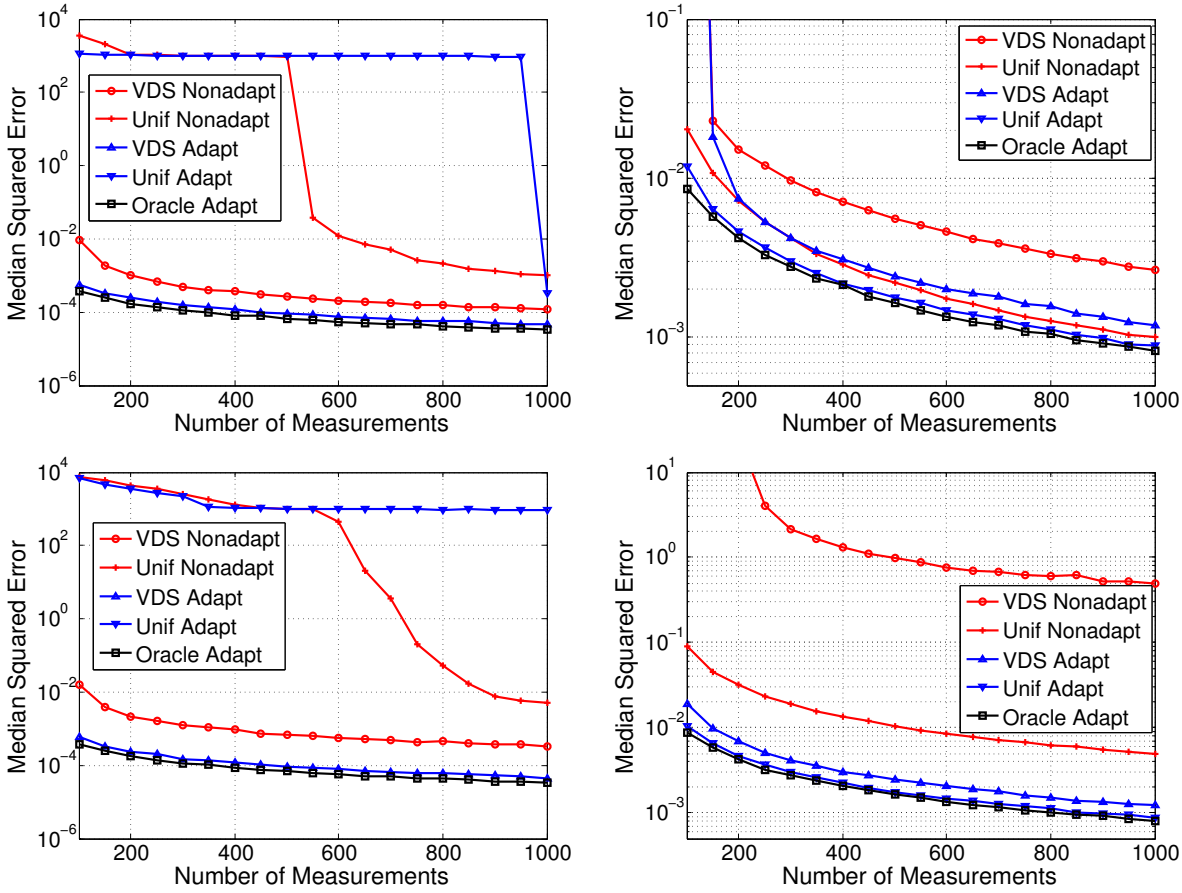


Figure 2.2: (Large measurement regime) The median squared error versus the number of measurements m when the nonzero locations of $\boldsymbol{\alpha}$ are selected on a sparse tree (left) or uniformly at random (right). The nonadaptive (red) and adaptive (blue) recovery is shown when either VDS or uniform sampling is used for the nonadaptive measurements, and CoSaMP (top) or ℓ_1-minimization (bottom) is used; the oracle adaptive (black) recovery is also included for comparison.

of 1-sparse signals.

2.4.1 Simulations

Here we present a practical implementation of adaptive sensing obtained via the relaxation (2.14) which we then compare with the results of traditional nonadaptive sensing. To implement (2.14), we use the Templates for First-Order Conic Solvers (TFOCS) software package [107, 108].

For our first two sets of experiments, we set $\mathscr{M}$ to be the ensemble of n measurements from the $n \times n$ DFT matrix $\mathbf{F}$. We define $\mathbf{x}$ to be a 10-sparse signal in the Haar wavelet basis (i.e., $\boldsymbol{\Psi} = \mathbf{H}^\star$) with the values on the support of $\boldsymbol{\alpha}$ distributed i.i.d as $\mathcal{N}(\sqrt{n}, 1)$ and

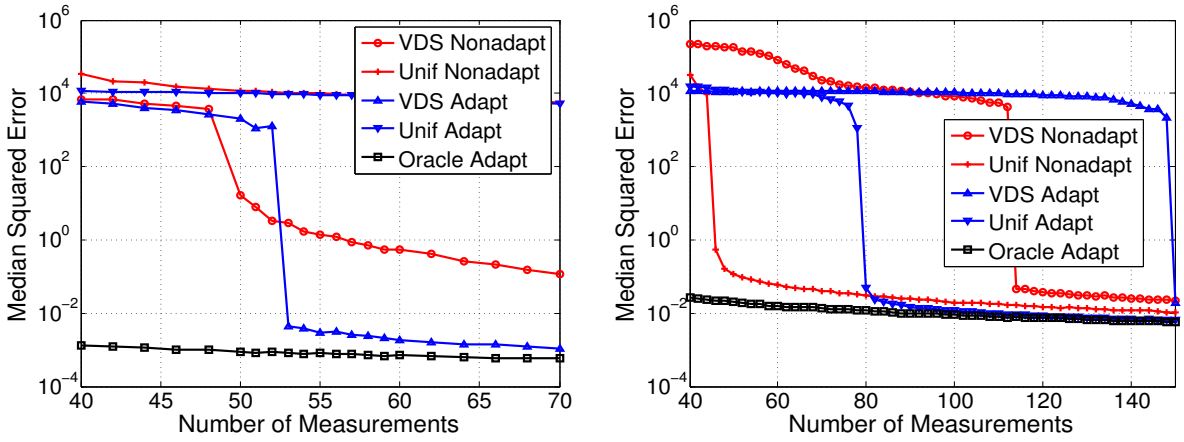


Figure 2.3: (Small measurement regime) The median squared error versus the number of measurements m when the nonzero locations of $\boldsymbol{\alpha}$ are selected on a sparse tree (left) or uniformly at random (right). The nonadaptive (red) and adaptive (blue) recovery is shown when either VDS or uniform sampling is used for the nonadaptive measurements, and CoSaMP is used; the oracle adaptive (black) recovery is also included for comparison.

the measurement noise $\mathbf{z}$ is distributed as i.i.d. $\mathcal{N}(0, 10^{-4})$.[6] Unless otherwise stated, the signal is of dimension $n = 1024$. We consider signals whose support is chosen uniformly at random, and also those whose support obeys a tree structure. Briefly, in the latter case the support is organized on a binary tree, plus an extra node at the top. The first scaling (or lowest frequency) coefficient has just one child; the second and further wavelet coefficients have two children each. This model is characteristic of natural images which tend to have inter-scale correlations (see [109, 110] for similar wavelet-tree constructions). An s-sparse support is filled by choosing the first scaling location, and then in each of the $s - 1$ remaining rounds, choosing one node randomly among the unfilled nodes which currently have a chosen parent.

Nonadaptive sensing. Due to the lack of incoherence between the DFT and Haar bases, it has been observed (and recently theoretically shown [111, 112]) that so-called *Variable-Density Sampling* (VDS) is often preferable to standard uniform random selection of DFT measurements. In VDS[7], sampling can be concentrated on the lower frequencies, producing

[6] We have found the adaptive procedure to be robust to the noise level, and compare similarly to the corresponding nonadaptive procedure even for larger noise levels.

[7] Following the experiments in [111], we also do not apply any preconditioning to the sensing matrix.

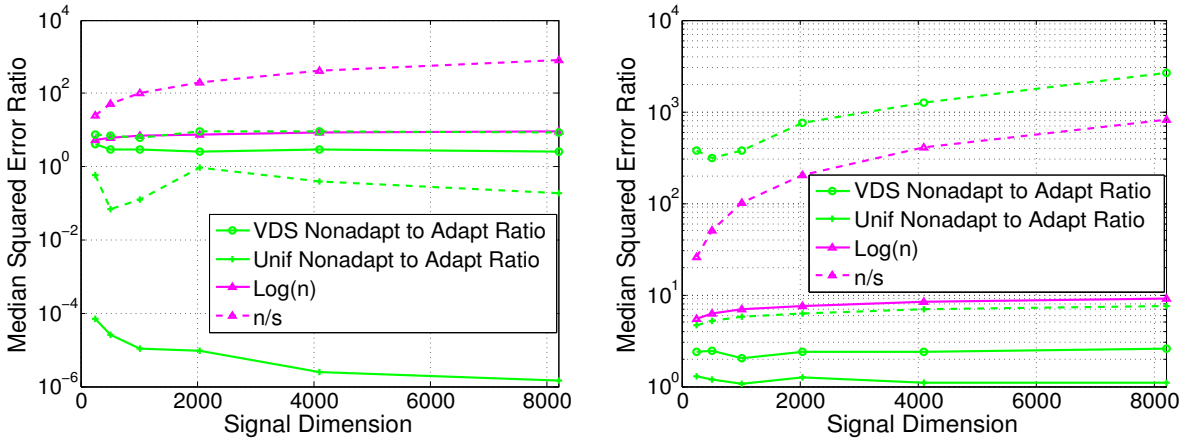


Figure 2.4: The ratio (green) of the nonadaptive median squared recovery error to the adaptive median squared recovery error versus the signal dimension n when the support locations of $\boldsymbol{\alpha}$ are selected on a sparse tree (left) or uniformly at random (right). The ratio is shown when either VDS or uniform sampling is used for the nonadaptive measurements, and CoSaMP (solid line) or ℓ_1-minimization (dashed line) is used. The curves for $\log n$ (solid magenta) and n/s (dashed magenta) are included for comparison.

superior recovery results. We test recovery using either ℓ_1-minimization [16, 113, 114] or the greedy pursuit CoSaMP [23].

Adaptive sensing. In the more realistic setting, we employ a simple strategy which uses $m/2$ nonadaptive measurements (using either VDS or uniform sampling) to construct an estimate of Λ. This is done by executing either ℓ_1-minimization (followed by thresholding) or CoSaMP. We then solve the relaxation (2.14) using this estimated support, and the remaining $m/2$ measurements are selected adaptively[8] using the distribution given by (2.15). To recover the signal, either ℓ_1-minimization or CoSaMP is again used to obtain an updated estimate $\hat{\Lambda}$ using all m measurements.[9] The final signal coefficient estimate is calculated as $\widehat{\boldsymbol{\alpha}}_{\hat{\Lambda}} = (\mathbf{F}'\mathbf{H}^*_{\hat{\Lambda}})^{\dagger}\mathbf{y}$, where $\mathbf{y}$ is the m-dimensional vector of measurements, $\mathbf{F}'$ is the $m \times n$ vector of DFT measurements selected, and $\mathbf{H}^*_{\hat{\Lambda}}$ is the $n \times 10$ submatrix of $\mathbf{H}^*$ restricted to the columns indexed by $\hat{\Lambda}$. One could alternatively use the signal estimate returned directly from the recovery algorithm, which we have observed to perform similarly to (or only

[8]Note that these $m/2$ *adaptively* selected measurements are only adapted to the first $m/2$ measurements, but are nonadaptive with respect to each other. That is, only one instance of adaptive measurement selection is being performed. Although in a different context, a similar two-stage approach is also taken in [99].

[9]Using dependent measurements is of course not justified theoretically, but we found unsurprisingly that using all m measurements gave better empirical results.

slightly worse than) our implemented method.

Oracle adaptive sensing. For sake of comparison, we also consider the case where the true support Λ of the signal is known a priori, and the measurements are selected as in the adaptive sensing case using this Λ. Recovery is then performed simply by applying the pseudoinverse: $\widehat{\boldsymbol{\alpha}}_\Lambda = (\mathbf{F}'\mathbf{H}_\Lambda^*)^\dagger \mathbf{y}$.

Figure 2.2 compares recovery results over 1000 trials for nonadaptive, adaptive, and oracle adaptive sensing versus the number of measurements m, where m ranges between 100 and 1000. We see that when the signal is supported on a tree, uniform sampling performs poorly for both nonadaptive and adaptive sensing, as might be expected. The performance of the uniform sampling methods can be understood via the empirical observation that in this case we require roughly 500 measurements before we can reliably estimate the support. When using the CoSaMP algorithm, the sudden improvement at $m \approx 500$ for uniform nonadaptive and at $m \approx 1000$ for uniform adaptive (which uses $m \approx 500$ measurements for support identification) corresponds to the threshold where more than half of the trials resulted in a correct support recovery. In contrast, sampling with VDS offers dramatic improvements for both nonadaptive and adaptive sensing with either reconstruction algorithm, with adaptive sensing performing almost as well as the oracle. In this case, VDS is already capturing much of the potential improvement offered by adaptivity because the energy of the signal is heavily biased towards the lower frequencies, although adaptivity still results in somewhat improved performance. In contrast to the tree-sparse case, when the signal support is selected randomly, uniform nonadaptive sampling actually performs better than VDS, whereas adaptive sensing performs similarly regardless of the type of nonadaptive measurements taken. Thus if one is not sure of the signal structure in general, adaptive sensing can offer improvements in either case. This flexibility represents one of the main advantages of adaptive sensing.

Figure 2.3 studies the same setting as Figure 2.2 when using CoSaMP for recovery, but focuses on the small measurement regime. These results illustrate that there are regions,

however narrow, where the nonadaptive method can succeed while the adaptive method fails. This is expected due to the nature of the adaptive scheme, where only $m/2$ measurements are utilized to identify a support estimate. At some point, the support can be sufficiently estimated with m, but not $m/2$, measurements. For tree-sparse signals, nonadaptive sensing with VDS measurements outperforms adaptive sensing with VDS nonadaptive measurements when $m \approx 50$. For uniformly sparse signals, we see this behavior even more clearly for both VDS and uniform sampling.

The results of our second simulation are shown in Figure 2.4, where we compare the ratio of nonadaptive to adaptive sensing recovery over 200 trials against the dimension n of the signal $\mathbf{x}$; the number of measurements used is always $m = 0.6n$ (rounding when necessary). We note that since the norms of $\boldsymbol{\alpha}$ and $\mathbf{z}$ both scale with n, the SNR remains roughly the same for all signal dimensions n. We observe similar results as Figure 2.2, demonstrating the behavior holds as a function of dimension.

In our last experiment, we evaluate our adaptive approach on real images. This scenario differs from previous experiments in two key aspects. First, the signal of interest is a two-dimensional (2D) image, not a one-dimensional vector, and thus we use 2D DFT measurements and a 2D discrete wavelet transform as the sparsity basis. Second, the image is not exactly sparse in any wavelet basis. Hence, when estimating the sparse support we introduce an additional (non-Gaussian) source of error, the contribution of the off-support wavelet coefficients. We note that the choice of the parameter s, which we have not attempted to optimize, can have an impact on signal reconstruction.

The image we use, `brain.mat`[10], is rescaled to be 64×64, and is shown in Figure 2.5. We use the Daubechies wavelet with 3 vanishing moments (D6) in a full 2D decomposition (i.e., $\log_2 64 = 6$ levels). We set the parameter $s = 1000$, which we again note was not tuned nor optimized. Additionally, we introduce white Gaussian noise at the level of $\sigma = 0.01$ to each measurement.

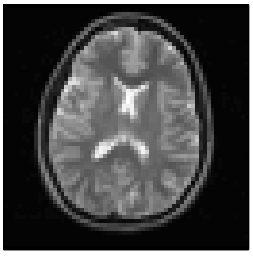
[10]Obtained from `http://www.eecs.berkeley.edu/~mlustig/CS.html`.

The experiment proceeds as follows: in the nonadaptive case, m measurements are taken according to VDS. The set of recovered wavelet coefficients are obtained using ℓ_1-minimization and the image is reconstructed using the inverse wavelet transform. Note that the output of ℓ_1-minimization is not necessarily exactly s-sparse. The assumed sparsity s guides our choice of the ℓ_2 error term constraint, but we did no thresholding afterwards. We evaluate performance by the median peak signal-to-noise ratio (PSNR) in dB over 50 trials.

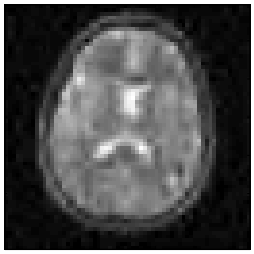
In the adaptive case, $m/2$ VDS measurements are taken as in the previous nonadaptive case. Then, via the ℓ_1-minimization reconstruction, we determine the estimated top s wavelet coefficients in each of 50 trials. We choose the trial with accuracy (in terms of the number of correctly identified top s wavelet coefficients) closest to the median accuracy. Utilizing the size s support estimate identified, we solve the relaxation (2.14) and select the remaining $m/2$ measurements adaptively. Finally, we recover the signal via ℓ_1-minimization, and, as before, reconstruct the final wavelet coefficients using the pseudoinverse. Again, we evaluate performance by the median PSNR over 50 trials of the adaptive measurement selection.

The adaptive and nonadaptive recovered images of the single trial with the closest to median PSNR performance using a total of $m = 3000$ measurements are given in Figure 2.5. Notice that the PSNR of the adaptive strategy is 28.02 dB, which exceeds the PSNR of 25.03 dB of the nonadaptive strategy. Visually, the adaptive strategy more closely resembles the original image. The median PSNR as the number of measurements is varied is shown in Figure 2.6. The plot shows that as the number of measurements reaches a certain level (roughly above 2000 measurements), the two-stage adaptive approach begins to exceed the method which is purely nonadaptive. Hence, as long as enough nonadaptive measurements are taken to obtain a sufficient support estimate, the adaptive procedure can improve image reconstruction quality.

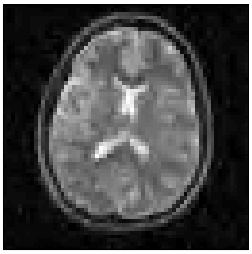
We note that adaptive approaches to medical imaging have also been studied using an alternative Bayesian model for sampling optimization [115–117]. The work [117] studies



Original



Nonadaptive,
PSNR = 25.03 dB.



Adaptive,
PSNR = 28.02 dB.

Figure 2.5: (Top) The 64×64 `brain.mat` image used in our medical imaging experiments. Reconstructed images with the closest to median PSNR among the 50 trials of nonadaptive (bottom left) and adaptive (bottom right) sensing with $m = 3000$ measurements.

the optimization of sequential sampling over stacks of neighboring image slices. In future work, it would be interesting to extend our proposed adaptive sampling scheme to this setting. Our method, however, is a framework for general adaptive sensing, not tuned specifically for medical imaging.

2.4.2 Analysis of the 1-sparse case

We now provide some analytical justification explaining why adaptive sensing can achieve a lower MSE than nonadaptive sensing for the Haar wavelet basis with DFT measurements, but show that the largest gains are realized for a small fraction of the possible signal support sets. We consider the simple case when $s = 1$ and the support is eventually known (either by oracle or by utilizing some method for estimation, as in the above experiments), and use this toy problem as motivational justification for the general setting. If we denote the s singular values of $\mathbf{F}'\mathbf{H}_\Lambda^*$ by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_s > 0$, then in general we want to minimize

$$\|(\mathbf{F}'\mathbf{H}_\Lambda^*)^\dagger\|_F^2 = \sum_{i=1}^{s} \frac{1}{\sigma_i^2}. \tag{2.19}$$

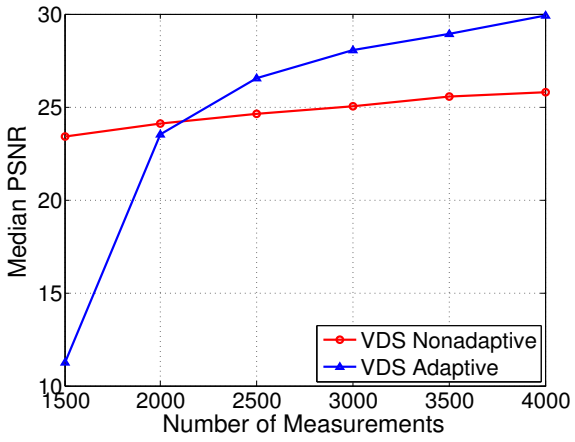


Figure 2.6: The median PSNR versus the number of measurements m over 50 trials of nonadaptive and adaptive sensing of the 64×64 `brain.mat` image.

When $s = 1$, (2.19) becomes $\|(\mathbf{F}'\mathbf{H}_\Lambda^*)^\dagger\|_F^2 = \frac{1}{\sigma_1^2}$. However, minimizing this quantity is the same as maximizing $\|\mathbf{F}'\mathbf{H}_\Lambda^*\|_F^2 = \sigma_1^2$. It is easy to see that $\|\mathbf{F}'\mathbf{H}_\Lambda^*\|_F^2$ is maximized when a measurement of $\mathbf{F}$ that is most correlated with $\mathbf{H}_\Lambda^*$ is chosen for every measurement in $\mathbf{F}'$. That is, if such a row can be identified, the best way to sense 1-sparse signals adaptively once the support is known is to simply repeat that measurement until the number of allotted measurements has been reached. Note that $\mathbf{F}'\mathbf{H}_\Lambda^*$ is $m \times 1$ in this case, and is still full rank when selecting the measurements in this way; thus, the theory leading to (2.18) still holds. In this setting, we can determine explicitly what the MSE looks like, and provide bounds on the MSE that depend on the support location of the 1-sparse signal. The result assuming a known support is provided in Theorem 2.4.1. The result in the more realistic context of adaptive sensing, where the first half of the measurements are selected nonadaptively, immediately follows and is provided in Corollary 2.4.5.

Theorem 2.4.1. *Denote by* $\mathbf{x} = \mathbf{H}^*\boldsymbol{\alpha}$ *the signal of interest, and suppose* $\mathbf{x}$ *becomes 1-sparse after applying the Haar wavelet transformation* $\mathbf{H}$ *(that is,* $\mathbf{H}\mathbf{x} = \boldsymbol{\alpha}$ *and* $\|\boldsymbol{\alpha}\|_0 = 1$*). Let* $\operatorname{supp}(\boldsymbol{\alpha}) = \Lambda$*, and suppose the support* Λ *is completely known. Suppose we measure repeatedly with a particular measurement from the* $n \times n$ *DFT* $\mathbf{F}$ *defined in (2.6); denote this measurement by* $\mathbf{f}_j$*, where* $j \in \{0, 1, \ldots, n-1\}$ *is some row index. Then, our observations*

are of the form

$$y_i = \langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \boldsymbol{\alpha}_\Lambda \rangle + z_i, \tag{2.20}$$

for $i = 1, \dots, m$, where the noise z_i are i.i.d. $N(0, \sigma^2)$. Then the MSE is given by

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 = \frac{1/m}{|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|^2} \sigma^2, \tag{2.21}$$

and is bounded by

$$\frac{\sigma^2}{m} \leq \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \frac{n\sigma^2}{2m}, \tag{2.22}$$

where the expectation is taken with respect to $\mathbf{z}$.

Note that in standard compressive sensing when we rely on the RIP, the DFT matrix $\mathbf{F}$ is normalized by $\frac{1}{\sqrt{m}}$ rather than $\frac{1}{\sqrt{n}}$. If we make this normalization, the bound in (2.22) becomes

$$\frac{\sigma^2}{n} \leq \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \frac{\sigma^2}{2}.$$

Including pre-conditioning and other scalings of course yields an analogous bound.

The proof of Theorem 2.4.1 relies on three lemmas that provide bounds for the term $|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|$ appearing in the MSE in (2.21). Specifically, one term of interest is

$$\min_{\Lambda \in \{0,\dots,n-1\}} \max_{j \in \{0,\dots,n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|. \tag{2.23}$$

The maximization over j corresponds to selecting the best DFT measurement $\mathbf{f}_j$, and the minimization accounts for the worst case signal (i.e., the worst case support Λ). On the other hand, we also want to obtain a value that represents the best case signal so we are also interested in

$$\max_{\Lambda \in \{0,\dots,n-1\}} \max_{j \in \{0,\dots,n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|. \tag{2.24}$$

Before proving Theorem 2.4.1, let us set some notation. The Haar wavelet transform matrix $\mathbf{H}$, defined in (2.16) and (2.17) consists of *blocks* of consecutive rows with the same

nonzero entry magnitudes. Let $1 \leq a \leq \log_2 n$ denote the block of $\mathbf{H}$, where $a = 1$ corresponds to the $\frac{n}{2}$ rows indexed by $j = \frac{n}{2}, \ldots, n-1$ (i.e., the "bottom" half of $\mathbf{H}$), $a = 2$ corresponds to the $\frac{n}{4}$ rows indexed by $j = \frac{n}{2} - \frac{n}{4}, \ldots, \frac{n}{2} - 1$, and so on. Similarly, for $\mathbf{H}^*$, instead of blocks of *rows*, we have blocks of *columns*; the block corresponding to $a = \log_2 n$ represents the lowest frequency wavelets, and the block corresponding to $a = 1$ represents the highest frequency wavelets.

Proof of Theorem 2.4.1. This proof requires the following three lemmas. Lemmas 2.4.2 and 2.4.3 are used to prove Lemma 2.4.4, and Lemma 2.4.4 is used to complete the proof of Theorem 2.4.1. The lemmas can be derived using elementary trigonometric bounds, and we omit the proofs here.

Lemma 2.4.2. *Fix $j \in \mathbb{Z}$ where $1 \leq j \leq n-1$ and let $a = 1, \ldots, \log_2 n$. Choose $k \in \mathbb{Z}$, $0 \leq k \leq n - \frac{2^a}{2}$. Then*

$$\left| \sum_{q=k}^{k+\frac{2^a}{2}-1} e^{-2\pi i j q/n} \right| = \sqrt{\frac{1 - \cos(\frac{2^a \pi j}{n})}{1 - \cos(\frac{2\pi j}{n})}}. \tag{2.25}$$

Lemma 2.4.3. *Let $\mathbf{f}_j$, $j \in \{0, \ldots, n-1\}$, be row j from the $n \times n$ DFT and let $\mathbf{H}^*_\Lambda$ be the inverse discrete Haar wavelet transform restricted to the column indexed by Λ. Let $a = 1, \ldots, \log_2 n$ denote the block of $\mathbf{H}^*$ and let $\Lambda \in \{1, 2, \ldots, n-1\}$, $|\Lambda| = 1$, be a column in the set corresponding to block a. Then,*

$$|\langle \mathbf{f}_j, \mathbf{H}^*_\Lambda \rangle| = \frac{1}{\sqrt{n 2^{a-1}}} \frac{1 - \cos(\frac{2^a \pi j}{n})}{\sqrt{1 - \cos(\frac{2\pi j}{n})}}, \tag{2.26}$$

where $j = 1, \ldots, n-1$. When $j = 0$,

$$|\langle \mathbf{f}_0, \mathbf{H}^*_\Lambda \rangle| = \begin{cases} 1 & \Lambda = 0 \\ 0 & \Lambda \in \{1, \ldots, n-1\}. \end{cases}$$

When $\Lambda = \{0\}$,

$$|\langle \mathbf{f}_j, \mathbf{H}_0^* \rangle| = \begin{cases} 1 & j = 0 \\ 0 & j = 1, 2, \ldots, n-1. \end{cases}$$

Lemma 2.4.4. *Let* $\mathbf{f}_j$, $j \in \{0, \ldots, n-1\}$, *be a row from the* $n \times n$ *DFT matrix* $\mathbf{F}$ *and let* $\mathbf{H}_\Lambda^*$ *be the inverse discrete Haar wavelet transform restricted to the column indexed by* Λ. *Let* $\Lambda \in \{0, 1, \ldots, n-1\}$ *so that* $|\Lambda| = 1$. *Then*

$$\min_{\Lambda \in \{0,\ldots,n-1\}} \max_{j \in \{0,\ldots,n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle| = \sqrt{\frac{2}{n}} \tag{2.27}$$

and

$$\max_{\Lambda \in \{0,\ldots,n-1\}} \max_{j \in \{0,\ldots,n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle| = 1. \tag{2.28}$$

Since $|\Lambda| = 1$, $\boldsymbol{\alpha}_\Lambda$ is just a scalar, so that the measurements (2.20) can be written as

$$y_i = \langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle \boldsymbol{\alpha}_\Lambda + z_i. \tag{2.29}$$

This can be concisely written as

$$\mathbf{y} = \mathbf{A}\boldsymbol{\alpha}_\Lambda + \mathbf{z}, \tag{2.30}$$

where $\mathbf{A}$ is the m-dimensional column vector with each entry equal to $\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle$. To estimate $\boldsymbol{\alpha}_\Lambda$, we apply $\mathbf{A}^\dagger$ to $\mathbf{y}$. In this case, $\mathbf{A}^\dagger$ is an m-dimensional row vector, with each entry equal to $\frac{1}{m\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle}$. Therefore,

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_\Lambda &= \mathbf{A}^\dagger \mathbf{y} = \mathbf{A}^\dagger (\mathbf{A}\boldsymbol{\alpha}_\Lambda + \mathbf{z}) \\ &= \begin{bmatrix} \frac{1}{m\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle} & \cdots & \frac{1}{m\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle} \end{bmatrix} \left(\begin{bmatrix} \langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle \\ \vdots \\ \langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle \end{bmatrix} \boldsymbol{\alpha}_\Lambda + \mathbf{z} \right) \\ &= \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{\alpha}_\Lambda + \frac{z_i}{\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle} \right) = \boldsymbol{\alpha}_\Lambda + \frac{1}{m} \sum_{i=1}^{m} \frac{z_i}{\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle}. \end{aligned}$$

Using this, and since $\sum_{i=1}^{m} z_i \sim \mathcal{N}(0, m\sigma^2)$, we find

$$\begin{aligned}
\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 &= \mathbb{E}\|\mathbf{H}^*(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\|_2^2 \\
&= \mathbb{E}\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2^2 = \mathbb{E}|\hat{\boldsymbol{\alpha}}_\Lambda - \boldsymbol{\alpha}_\Lambda|^2 \\
&= \mathbb{E}\left|\boldsymbol{\alpha}_\Lambda - \left(\boldsymbol{\alpha}_\Lambda + \frac{1}{m}\sum_{i=1}^{m} \frac{z_i}{\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle}\right)\right|^2 \\
&= \mathbb{E}\left|\frac{1}{m}\sum_{i=1}^{m} \frac{z_i}{\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle}\right|^2 \\
&= \left(\frac{1/m}{|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|}\right)^2 \mathbb{E}\left|\sum_{i=1}^{m} z_i\right|^2 \\
&= \left(\frac{1/m}{|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|}\right)^2 m\sigma^2 = \frac{1/m}{|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|^2}\sigma^2. \qquad (2.31)
\end{aligned}$$

Applying the bounds from Lemma 2.4.4 to (2.31), we arrive at

$$\frac{\sigma^2}{m} \leq \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \frac{n\sigma^2}{2m},$$

as desired. □

Corollary 2.4.5. *Suppose* $\mathbf{x} = \mathbf{H}^\star\boldsymbol{\alpha}$ *is 1-sparse. Suppose after* $\frac{m}{2}$ *nonadaptive DFT measurements, the support* Λ *is correctly identified. For the remaining* $\frac{m}{2}$ *DFT measurements, we measure repeatedly with a particular measurement from the* $n \times n$ *DFT* $\mathbf{F}$*; denote this measurement by* $\mathbf{f}_j$*, where* $j \in \{0, 1, \ldots, n-1\}$ *is some row index. Then our observations are of the form (2.20) for* $i = \frac{m}{2} + 1, \ldots, m$*, where the noise* z_i *are i.i.d.* $N(0, \sigma^2)$*. Then the MSE is given by*

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 = \frac{2/m}{|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|^2}\sigma^2, \qquad (2.32)$$

and is bounded by

$$\frac{2\sigma^2}{m} \leq \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \frac{n\sigma^2}{m}, \qquad (2.33)$$

where the expectation is taken with respect to $\mathbf{z}$*.*

The upper bound on $\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2$ in Corollary 2.4.5 is precisely the lower bound from

Theorem 2.2.1 when $s = 1$. This means that there is indeed some room for improvement with adaptive sensing when the sparsity basis is the Haar wavelet transform rather than the canonical basis. Corollary 2.4.5 shows that the performance of adaptive sensing, in terms of the MSE, depends on the support location of 1-sparse signals. The best adaptive recovery is possible when the support is located on the lowest wavelet frequency ($\Lambda = \{0\}$, or the first Haar wavelet coefficient) while the worst recovery occurs when the support is located on any of the higher wavelet frequencies in block $a = 1$ (the latter half of the Haar wavelet coefficients). This of course matches the intuition based on the correlations in these two bases. This suggests that structured signals such as those that are tree-sparse will benefit more from adaptivity than signals that have a uniformly distributed support.

In light of the discrepancy between (2.27) and (2.28), one wishes to know in some sense, what fraction of signals allow for recovery more like one versus the other. Figure 2.7 shows how $\max_{j\in\{0,\dots,n-1\}} |\langle \mathbf{f}_j, \mathbf{H}^*_\Lambda \rangle|$ varies by maximizing $\max_{j\in\{0,\dots,n-1\}} |\langle \mathbf{f}_j, \mathbf{H}^*_\Lambda \rangle|$ over Λ while successively removing blocks from $\mathbf{H}^*$. Using our notation for blocks, the blocks of $\mathbf{H}^*$ are removed in the following (top-down) order: $\log_2(n), \log_2(n) - 1, \dots, 1$. Then, we plot the value of $\max_{j\in\{0,\dots,n-1\}} |\langle \mathbf{f}_j, \mathbf{H}^*_\Lambda \rangle|$ for the remaining submatrix of $\mathbf{H}^*$. Hence, we see that the MSE is higher for signals supported on higher wavelet frequencies, and the upper bound of (2.33) is achieved by exactly half of the possible signal support sets, whereas the lower bound of (2.33) is achieved by exactly one of the possible signal supports sets (i.e., $\Lambda = \{0\}$). Fortunately, the support of natural images tends to be concentrated on lower-frequency wavelet coefficients [118].

2.5 Discussion

Adaptive sensing has tremendous potential to improve the accuracy of sparse recovery in a variety of settings. However, in many practical applications one does not have the freedom to choose arbitrary measurement vectors, but instead must choose from a specified pool of measurements. One example of particular interest is the setting where measurements must be taken from the Fourier ensemble, as is the case in many medical imaging applications. In

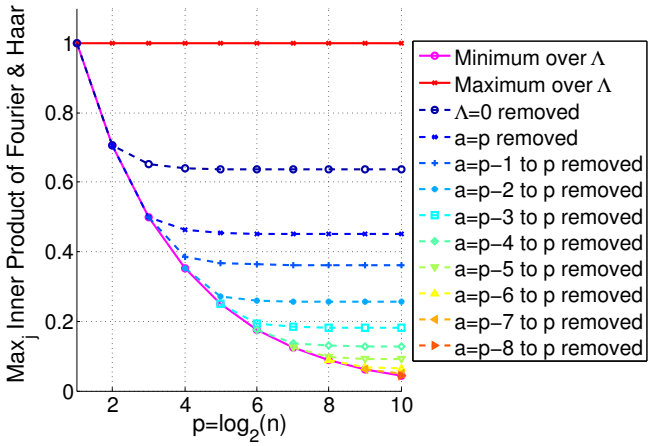


Figure 2.7: The value of $\max_{j\in\{0,\dots,n-1\}} |\langle \mathbf{f}_j, \mathbf{H}^*_\Lambda \rangle|$ is displayed against the (log of the) signal dimension $n = 2^p$. The solid magenta curve shows the optimization when minimizing over all possible supports $\Lambda \in \{0, \dots, n-1\}$, given by (2.27). The solid red curve shows the opposite optimization when maximizing over all possible supports $\Lambda \in \{0, \dots, n-1\}$, given by (2.28). The remaining dashed curves show the optimization when maximizing over all supports Λ except those in the blocks indicated.

this chapter we established fundamental limitations on the improvements offered by adaptivity in this setting for certain sparsity bases. On the other hand, we argued that for other sparsity bases (such as the Haar wavelet basis) the role of adaptivity in the constrained setting is much less straightforward. We developed a sampling scheme which uses a simple optimization procedure to select measurements adapted to the signal support. This scheme results in significant improvements once an accurate estimate of the support is obtained, which in practice can be achieved by first dedicating a portion of the measurements to support estimation. Though this approach is not necessarily provably optimal, it nonetheless demonstrates the potential of adaptive sensing in the constrained setting. We believe future work in this area can further the understanding of both the limitations of this approach as well as the potential benefits.

CHAPTER 3

LOCALIZATION VIA PAIRED COMPARISONS

Suppose that we wish to estimate a vector $\mathbf{x}$ from a set of binary paired comparisons of the form "$\mathbf{x}$ is closer to $\mathbf{p}$ than to $\mathbf{q}$" for various choices of vectors $\mathbf{p}$ and $\mathbf{q}$. The problem of estimating $\mathbf{x}$ from this type of observation arises in a variety of contexts, including nonmetric multidimensional scaling, "unfolding," and ranking problems, often because it provides a powerful and flexible model of preference. We describe theoretical bounds for how well we can expect to estimate $\mathbf{x}$ under a randomized model for $\mathbf{p}$ and $\mathbf{q}$. We also present results for the case where the comparisons are noisy and subject to some degree of error. Additionally, we show that under a randomized model for $\mathbf{p}$ and $\mathbf{q}$, a suitable number of binary paired comparisons yield a stable embedding of the space of target vectors. Finally, we show that we can achieve significant gains by adaptively altering the distribution for choosing $\mathbf{p}$ and $\mathbf{q}$.

3.1 Introduction

3.1.1 The localization problem

In this chapter we consider the problem of determining the location of a point in Euclidean space based on distance comparisons to a set of known points, where our observations are nonmetric. In particular, let $\mathbf{x} \in \mathbb{R}^n$ be the true position of the point that we are trying to estimate, and let $(\mathbf{p}_1, \mathbf{q}_1), \ldots, (\mathbf{p}_m, \mathbf{q}_m)$ be pairs of "landmark" points in $\mathbb{R}^n$ which we assume to be known *a priori*. Rather than directly observing the raw distances from $\mathbf{x}$, i.e., $\|\mathbf{x} - \mathbf{p}_i\|$ and $\|\mathbf{x} - \mathbf{q}_i\|$, we instead obtain only paired comparisons of the form $\|\mathbf{x} - \mathbf{p}_i\| < \|\mathbf{x} - \mathbf{q}_i\|$. Our goal is to estimate $\mathbf{x}$ from a set of such inequalities. Nonmetric observations of this type arise in numerous applications and have seen considerable interest in recent literature e.g., [122–125]. These methods are often applied in situations where we have a collection

Material in this section is work with Mark Davenport and is available in preprint [119], being prepared for publication. It has lead to publications [120, 121].

of items and hypothesize that it is possible to embed the items in $\mathbb{R}^n$ in such a way that the Euclidean distance between points corresponds to their "dissimilarity," with small distances corresponding to similar items. Here, we focus on the sub-problem of adding a new point to a known (or previously learned) configuration of landmark points.

As a motivating example, we consider the problem of estimating a user's preferences from limited response data. This is useful, for instance, in recommender systems, information retrieval, targeted advertising, and psychological studies. A common and intuitively appealing way to model preferences is via the *ideal point model*, which supposes preference for a particular item varies inversely with Euclidean distance in a feature space [126]. We assume that the items to be rated are represented by points $\mathbf{p}_i$ and $\mathbf{q}_i$ in an n-dimensional Euclidean space. A user's preference is modeled as an additional point $\mathbf{x}$ in this space (called the individual's "ideal point"). This represents a hypothetical "perfect" item satisfying all of the user's criteria for evaluating items.

Using response data consisting of paired comparisons between items (e.g., "user $\mathbf{x}$ prefers item $\mathbf{p}_i$ to item $\mathbf{q}_i$") is a natural approach when dealing with human subjects since it avoids requiring people to assign precise numerical scores to different items (which is generally a quite difficult task, especially when preferences may depend on multiple factors [127]). In contrast, human subjects often find pairwise judgements much easier to make [128]. Data consisting of paired comparisons is often generated implicitly in contexts where the user has the option to act on two (or more) alternatives; for instance they may choose to watch a particular movie, or click a particular advertisement, out of those displayed to them [129]. In such contexts, the "true distances" in the ideal point model's preference space are generally inaccessible directly, but it is nevertheless still possible to obtain an estimate of a user's ideal point.

3.1.2 Main results

The fundamental question which interests us in this chapter is how many comparisons we need (and how should we choose them) to estimate $\mathbf{x}$ to a desired degree of accuracy. Thus,

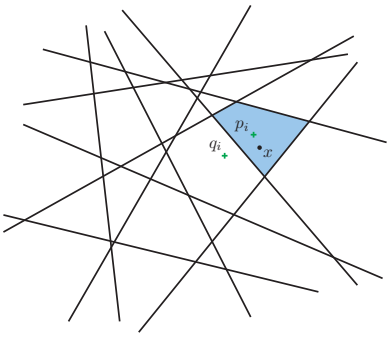


Figure 3.1: An illustration of the localization problem from paired comparisons. The information that $\mathbf{x}$ is closer to $\mathbf{p}_i$ than $\mathbf{q}_i$ tells us which side of a hyperplane $\mathbf{x}$ lies. Through many such comparisons we can hope to localize $\mathbf{x}$ to a high degree of accuracy.

we consider the case where we are given an existing embedding of the items (as in a mature recommender system) and focus on the on-line problem of locating a single new user from their feedback (consisting of binary data generated from paired comparisons). The item embedding could be generated using various methods, such as multidimensional scaling applied to a set of item features, or even using the results of previous paired comparisons via an approach like that in [130]. Given such an embedding of ℓ items, there are a total of $\binom{\ell}{2} = \Theta(\ell^2)$ possible paired comparisons. Clearly, in a system with thousands (or more) items, it will be prohibitive to acquire this many comparisons as a typical user will likely only provide comparisons for a handful of items. Fortunately, in general we can expect that many, if not most, of the possible comparisons are actually redundant. For example, of the comparisons illustrated in Figure 3.1, all but four are redundant and – at least in the absence of noise – add no additional information.

Any precise answer to this question would depend on the underlying geometry of the item embedding. Each comparison essentially divides $\mathbb{R}^n$ in two, indicating on which side of a hyperplane $\mathbf{x}$ lies, and some arrangements of hyperplanes will yield better tessellations of the preference space than others. Thus, to gain some intuition on this problem without reference to the geometry of a particular embedding, we will instead consider a probabilistic model where the items are generated at random from a particular distribution. In this case we show that under certain natural assumptions on the distribution, it is possible to estimate

the location of any $\mathbf{x}$ to within an error of ϵ using a number of comparisons which, up to log factors, is proportional to n/ϵ. This is essentially optimal, so that no set of comparisons can provide a uniform guarantee with significantly fewer comparisons. We then describe several stability and robustness guarantees for various settings in which the comparisons are subject to noise or errors. Finally, we then describe a simple extension to an *adaptive* scheme where we adaptively select the comparisons (manifested here in adaptively altering the mean and variance of the distribution generating the items) to substantially reduce the required number of comparisons.

3.1.3 Related work

It is important to note that the ideal point model, while similar, is distinct from the low-rank model used in *matrix completion* [131, 132]. Although both models suppose user choices are guided by a number of attributes, the ideal point model leads to preferences that are *non-monotonic* functions of those attributes. The ideal point model suggests that each feature has an ideal level; too much of a feature can be just as undesirable as too little. It is not possible to obtain this kind of performance with a traditional low-rank model, though if points are limited to the sphere, then the ideal point model can duplicate the performance of a low-rank factorization. There is also empirical evidence that the ideal point model captures behavior more accurately than factorization based approaches do [133, 134].

There is a large body of work that studies the problem of learning to rank items from various sources of data, including paired comparisons of the sort we consider in this chapter. See, for example, [135–137] and references therein. We first note that in most work on rankings, the central focus is on learning a correct rank-ordered list for a particular user, without providing any guarantees on recovering a correct parameterization for the user's preferences as we do here. While these two problems are related, there are natural settings where it might be desirable to guarantee an accurate recovery of the underlying parameterization ($\mathbf{x}$ in our model). For example, one could exploit these guarantees in the context of an iterative algorithm for nonmetric multidimensional scaling which aims to refine the un-

derlying embedding by updating each user and item one at a time (e.g., see [138]), in which case an understanding of the error in the estimate of $\mathbf{x}$ is crucial. Moreover, we believe that our approach provides an interesting alternative perspective as it yields natural robustness guarantees and suggests simple adaptive schemes.

Perhaps most closely related to our work is that of [135], which examines the problem of learning a rank ordering using the same ideal point model considered in this chapter. The message in this work is broadly consistent with ours, in that the number of comparisons required should scale with the dimension of the preference space (not the total number of items) and can be significantly improved via a clever adaptive scheme. However, this work does not bound the estimation error in terms of the Euclidean distance, which is our central concern. [136] also incorporates adaptivity, but seeks to embed a set of points in Euclidean space (as opposed to a single user's ideal point) and relies on paired comparisons involving three arbitrarily selected points (rather than a user's ideal point and two items).

Also closely related is the work in [139–141] which consider paired comparisons and more general ordinal measurements in the similar (but as discussed above, subtly different) context of low-rank factorizations. Finally, while seemingly unrelated, we note that our work builds on the growing body of literature of 1-bit compressive sensing. In particular, our results are largely inspired by those in [72, 74], and borrow techniques from [69] in the proofs of some of our main results. Note that in this work we extend preliminary results first presented in [120, 121].

3.2 A randomized observation model

For the moment we will consider the "noise-free" setting where each comparison between $\mathbf{x}$ and $\mathbf{q}_i$ versus $\mathbf{p}_i$ results in assigning the point which is truly closest to $\mathbf{x}$ with probability 1. In this case we can represent the observed comparisons mathematically by letting $\mathscr{A}_i(\mathbf{x})$

denote the i^{th} observation, which consists of comparisons between $\mathbf{p}_i$ and $\mathbf{q}_i$, and setting

$$\mathscr{A}_i(\mathbf{x}) := \operatorname{sign}\left(\|\mathbf{x}-\mathbf{q}_i\|^2 - \|\mathbf{x}-\mathbf{p}_i\|^2\right) = \begin{cases} +1 & \text{if } \mathbf{x} \text{ is closer to } \mathbf{p}_i \\ -1 & \text{if } \mathbf{x} \text{ is closer to } \mathbf{q}_i. \end{cases} \tag{3.1}$$

We will also use $\mathscr{A}(\mathbf{x}) := [\mathscr{A}_1(\mathbf{x}), \cdots, \mathscr{A}_m(\mathbf{x})]^T$ to denote the vector of all observations resulting from m comparisons. Note that since

$$\|\mathbf{x}-\mathbf{q}_i\|^2 - \|\mathbf{x}-\mathbf{p}_i\|^2 = 2(\mathbf{p}_i-\mathbf{q}_i)^T\mathbf{x} + \|\mathbf{q}_i\|^2 - \|\mathbf{p}_i\|^2,$$

if we set $\bar{\mathbf{a}}_i = (\mathbf{p}_i - \mathbf{q}_i)$ and $\bar{\tau}_i = \frac{1}{2}(\|\mathbf{p}_i\|^2 - \|\mathbf{q}_i\|^2)$, then we can re-write our observation model as

$$\mathscr{A}_i(\mathbf{x}) = \operatorname{sign}\left(2\bar{\mathbf{a}}_i^T\mathbf{x} - 2\bar{\tau}_i\right) = \operatorname{sign}\left(\bar{\mathbf{a}}_i^T\mathbf{x} - \bar{\tau}_i\right). \tag{3.2}$$

This is reminiscent of the standard setup in one-bit compressive sensing (with dithers) [72, 74] with the important differences that: *(i)* we have not yet made any kind of sparsity or other structural assumption on $\mathbf{x}$ and, *(ii)* the "dithers" $\bar{\tau}_i$, at least in this formulation, are dependent on the $\bar{\mathbf{a}}_i$, which results in difficulty applying standard results from this theory to the present setting.

However, many of the techniques from this literature will nevertheless be helpful in analyzing this problem. To see this, we consider a randomized observation model where the pairs $(\mathbf{p}_i, \mathbf{q}_i)$ are chosen independently with i.i.d. entries drawn according to a normal distribution, i.e., $\mathbf{p}_i, \mathbf{q}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. In this case, we have that the entries of our sensing vectors are i.i.d. with $\bar{a}_i(j) \sim \mathcal{N}(0, 2\sigma^2)$. Moreover, if we define $\mathbf{b}_i = \mathbf{p}_i + \mathbf{q}_i$, then we also have that $\mathbf{b}_i \sim \mathcal{N}(0, 2\sigma^2\mathbf{I})$, and

$$\begin{aligned} \frac{1}{2}\bar{\mathbf{a}}_i^T\mathbf{b}_i &= \frac{1}{2}\sum_j (\mathbf{p}_i(j) - \mathbf{q}_i(j))(\mathbf{p}_i(j) + \mathbf{q}_i(j)) \\ &= \frac{1}{2}\sum_j \mathbf{p}_i(j)^2 - \mathbf{q}_i(j)^2 = \frac{1}{2}(\|\mathbf{p}_i\|^2 - \|\mathbf{q}_i\|^2) = \bar{\tau}_i. \end{aligned}$$

Note that while $\bar{\tau}_i = \frac{1}{2}\bar{\mathbf{a}}_i^T\mathbf{b}_i$ is clearly dependent on $\bar{\mathbf{a}}_i$, we do have that $\bar{\mathbf{a}}_i$ and $\mathbf{b}_i$ are inde-

pendent.

To simplify, we re-normalize by dividing by $\|\bar{\mathbf{a}}_i\|$, i.e., setting $\mathbf{a}_i := \bar{\mathbf{a}}_i/\|\bar{\mathbf{a}}_i\|$ and $\tau_i := \bar{\tau}_i/\|\bar{\mathbf{a}}_i\|$, in which case we can write

$$\mathcal{A}_i(\mathbf{x}) = \operatorname{sign}\left(\mathbf{a}_i^T\mathbf{x} - \tau_i\right). \tag{3.3}$$

It is easy to see that $\mathbf{a}_i$ is distributed uniformly on the sphere $\mathbb{S}^{n-1} = \{\mathbf{a} \in \mathbb{R}^n : \|\mathbf{a}\| = 1\}$. Note that throughout our analysis we will exploit the fact that $\mathbf{a}_i$ is uniform on $\mathbb{S}^{n-1}$ and will let ν denote the uniform measure on the sphere. Note also that

$$\tau_i = \frac{1}{2}\mathbf{a}_i^T\mathbf{b}_i.$$

Since $\bar{\mathbf{a}}_i$ and $\mathbf{b}_i$ are independent, $\mathbf{a}_i$ and $\mathbf{b}_i$ are also independent. Moreover, for any unit-vector $\mathbf{a}_i$, if $\mathbf{b}_i \sim \mathcal{N}(0, 2\sigma^2\mathbf{I})$ then $\mathbf{a}_i^T\mathbf{b}_i \sim \mathcal{N}(0, 2\sigma^2)$. Thus, we must have $\tau_i \sim \mathcal{N}(0, \sigma^2/2)$, independent of $\mathbf{a}_i$, which is the key insight that enables the analysis below.

3.3 Guarantees in the noise-free setting

We now state our main result concerning localization under the noise-free random model from Section 3.2. Let $\mathbb{B}_R^n$ denote the n-dimensional, radius R Euclidean ball.

Theorem 3.3.1. *Let $\epsilon, \eta > 0$ be given. Let $\mathcal{A}_i(\cdot)$ be defined as in* (3.1)*, and suppose that m pairs $\{(\mathbf{p}_i, \mathbf{q}_i)\}_{i=1}^m$ are generated by drawing each $\mathbf{p}_i$ and $\mathbf{q}_i$ independently from $\mathcal{N}(0, \sigma^2 I)$ where $\sigma^2 = 2R^2/n$. There exists a constant C such that if*

$$m \geq C\frac{R}{\epsilon}\left(n\log\frac{R\sqrt{n}}{\epsilon} + \log\frac{1}{\eta}\right), \tag{3.4}$$

then with probability at least $1-\eta$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{B}_R^n$ such that $\mathcal{A}(\mathbf{x}) = \mathcal{A}(\mathbf{y})$,

$$\|\mathbf{x} - \mathbf{y}\| \leq \epsilon.$$

The result follows from applying Lemma 3.3.2 below to pairs of points in a covering set of $\mathbb{B}_R^n$. The key message of this theorem is that if one chooses the variance σ^2 of the

distribution generating the items appropriately, then it is possible to estimate $\mathbf{x}$ to within ϵ using a number of comparisons that is nearly linear in n/ϵ. A natural question is what would happen with a different choice of σ^2. In fact, this assumption is critical—if σ^2 is substantially smaller the bound quickly becomes vacuous, and as σ^2 grows much past R^2/n the bound begins to become steadily worse.[1] As we will see in Section 3.6, this is in fact observed in practice. It should also be somewhat intuitive: if σ^2 is too small, then nearly all the hyperplanes induced by the comparisons will pass very close to the origin, so that accurate estimation of even $\|\mathbf{x}\|$ becomes impossible. On the other hand, if σ^2 is too large, then an increasing number of these hyperplanes will not even intersect the ball of radius R in which $\mathbf{x}$ is presumed to lie, thus yielding no new information.

Lemma 3.3.2. *Let $\mathbf{w}, \mathbf{z} \in \mathbb{B}_R^n$ be distinct and fixed, and let $\delta > 0$ be given. Define*

$$B_\delta(\mathbf{w}) := \{\mathbf{u} \in \mathbb{B}_R^n : \|\mathbf{u} - \mathbf{w}\| \leq \delta\}.$$

Let $\mathcal{A}_i$ be defined as in Theorem 3.3.1. Denote by P_{sep} the probability that $B_\delta(\mathbf{w})$ and $B_\delta(\mathbf{z})$ are separated by hyperplane i, i.e.,

$$P_{\text{sep}} := \mathbb{P}\left[\forall \mathbf{u} \in B_\delta(\mathbf{w}), \forall \mathbf{v} \in B_\delta(\mathbf{z}) : \mathcal{A}_i(\mathbf{u}) \neq \mathcal{A}_i(\mathbf{v})\right].$$

For any $\epsilon_0 \leq \|\mathbf{w} - \mathbf{z}\|$ we have

$$P_{\text{sep}} \geq \frac{\epsilon_0 - \delta\sqrt{2n}}{22\sqrt{\pi}e^{5/2}R}.$$

Proof. Let $\epsilon = \|\mathbf{w} - \mathbf{z}\|$. Here, we denote the normal vector and threshold of hyperplane i

[1] We note that it is possible to try to optimize σ^2 by setting $\sigma^2 = cR^2/n$ for some constant c and then selecting c so as to minimize the constant C in (3.4). We believe this would yield limited insight since, in order to obtain a result which is valid uniformly for all possible n, we use certain bounds which for general n can be somewhat loose and would skew the resulting c. We instead simply select $c = 2$ for simplicity in our analysis (as it results in $\tau_i \sim \mathcal{N}(0, R^2/n)$) and because it aligns well with simulations.

by $\mathbf{a}$ and τ respectively. It is easy to show that P_{sep} can be expressed as

$$\begin{aligned} P_{\text{sep}} &= \mathbb{P}\left[\mathbf{a}^T\mathbf{z} + \delta \le \tau \le \mathbf{a}^T\mathbf{w} - \delta \text{ or } \mathbf{a}^T\mathbf{w} + \delta \le \tau \le \mathbf{a}^T\mathbf{z} - \delta\right] \\ &= 2\,\mathbb{P}\left[\mathbf{a}^T\mathbf{z} + \delta \le \tau \le \mathbf{a}^T\mathbf{w} - \delta\right], \end{aligned} \tag{3.5}$$

where the second equality follows from the symmetry of the distributions of $\mathbf{a}$ and τ.

Define $C_\alpha := \{\mathbf{a} \in \mathbb{S}^{n-1} : \mathbf{a}^T(\mathbf{w} - \mathbf{z}) \ge \alpha\}$. Note that the probability in (3.5) is zero unless $\mathbf{a} \in C_{2\delta}$. Thus, recalling that $\tau_i \sim \mathcal{N}(0, \sigma^2/2)$ we have

$$\begin{aligned} P_{\text{sep}} &= 2\int_{C_{2\delta}} \left|\Phi\left(\frac{\mathbf{a}^T\mathbf{w} - \delta}{\sigma/\sqrt{2}}\right) - \Phi\left(\frac{\mathbf{a}^T\mathbf{z} + \delta}{\sigma/\sqrt{2}}\right)\right| \nu(\mathrm{d}\mathbf{a}) \\ &\ge 2\int_{C'} \left|\Phi\left(\frac{\mathbf{a}^T\mathbf{w} - \delta}{\sigma/\sqrt{2}}\right) - \Phi\left(\frac{\mathbf{a}^T\mathbf{z} + \delta}{\sigma/\sqrt{2}}\right)\right| \nu(\mathrm{d}\mathbf{a}) \end{aligned} \tag{3.6}$$

for any $C' \subseteq C_{2\delta}$. To obtain a lower bound on (3.6), we will consider a carefully chosen subset $C' \subseteq C_{2\delta}$ and then simply multiply the area of C' by the minimum value γ of the integrand over that set, yielding a bound of the form

$$P_{\text{sep}} \ge 2\gamma\nu(C').$$

We construct the set C' as follows. Let $W := \{\mathbf{a} : \mathbf{a}^T\mathbf{w} \le \xi/\sqrt{n}\|\mathbf{w}\|\}$, $Z := \{\mathbf{a} : \mathbf{a}^T\mathbf{z} \ge -\xi/\sqrt{n}\|\mathbf{z}\|\}$, and set $C' := C_\alpha \cap W \cap Z$ for some $\alpha \ge 2\delta$. Note that for any $\mathbf{a} \in C'$, since $\mathbf{a}^T(\mathbf{w} - \mathbf{z}) \ge \alpha \ge 2\delta$, we have $-R\xi/\sqrt{n} \le \mathbf{a}^T\mathbf{z} + \delta \le \mathbf{a}^T\mathbf{w} - \delta \le R\xi/\sqrt{n}$. Thus, by Lemma 3.8.1,

$$\gamma = \inf_{\mathbf{a}\in C'} \left|\Phi\left(\frac{\mathbf{a}^T\mathbf{w} - \delta}{\sigma/\sqrt{2}}\right) - \Phi\left(\frac{\mathbf{a}^T\mathbf{z} + \delta}{\sigma/\sqrt{2}}\right)\right| \ge \frac{\sqrt{2}}{\sigma}(\alpha - 2\delta)\phi\Big(\frac{\sqrt{2}R\xi}{\sigma\sqrt{n}}\Big).$$

Recall by assumption we have that $\sigma = \sqrt{2}R/\sqrt{n}$, thus we obtain by setting $\xi = \sqrt{5}$,

$$\gamma \ge \frac{\sqrt{n}}{R}(\alpha - 2\delta)\phi(\xi) = \frac{\sqrt{n}(\alpha - 2\delta)}{\sqrt{2\pi}e^{5/2}R}. \tag{3.7}$$

Next note that $C' = C_\alpha \cap W \cap Z = C_\alpha \setminus W^c \setminus Z^c$ is a difference of a set of hyperspherical caps. To obtain a lower bound on $\nu(C')$ we use the upper and lower bounds on the measure

of hyperspherical caps given in Lemma 2.1 of [142].

Case $n \geq 6$**.—** Provided that $\alpha/\epsilon < \sqrt{2/n}$ we can bound $\nu(C')$ as

$$\nu(C') \geq \nu(C_\alpha) - \nu(W^c) - \nu(Z^c) \geq \frac{1}{12} - 2\frac{1}{2\xi}(1-\xi^2/n)^{(n-1)/2} \geq \frac{1}{12} - \frac{1}{\sqrt{5}e^{5/2}},$$

where the last inequality follows from the fact that $(1-x/n)^{n-1} \leq e^{-x}$ for $n \geq x \geq 2$. Combining this with lower estimate (3.7),

$$P_{\text{sep}} \geq 2\gamma\nu(C') \geq 2\frac{\sqrt{n}(\alpha-2\delta)}{\sqrt{2\pi}e^2 R}\frac{1-12e^{-5/2}/\sqrt{5}}{12}.$$

Setting $\alpha = \delta + \epsilon/\sqrt{2n}$, since $1 - 12e^{-5/2}/\sqrt{5} > 5/9$, we have that

$$P_{\text{sep}} \geq \frac{2\sqrt{n}(\epsilon/\sqrt{2n}-\delta)(1-12e^{-5/2}/\sqrt{5})}{12\sqrt{2\pi}e^{5/2}R} \geq \frac{\epsilon - \delta\sqrt{2n}}{22\sqrt{\pi}e^{5/2}R}.$$

Note that this bound holds under the assumption that $\alpha/\epsilon < \sqrt{2/n}$, which for our choice of α is equivalent to the assumption that $\epsilon > \delta\sqrt{2n}$. However, this bound also holds trivially for all $\epsilon \leq \delta\sqrt{2n}$, and thus in fact holds for all $\epsilon \geq 0$.

Case $n \leq 5$**.—** In this case, note that $\xi/\sqrt{n} \geq 1$, so the sets W and Z are the entire sphere. Hence, $\nu(W^c) = \nu(Z^c) = 0$ and $\nu(C') = \nu(C_\alpha) \geq \frac{1}{12}$. Thus,

$$P_{\text{sep}} \geq 2\gamma\nu(C') \geq \frac{\epsilon - \delta\sqrt{2n}}{12\sqrt{\pi}e^{5/2}R}.$$

We obtain the stated lemma by noting $\epsilon_0 \leq \epsilon$. □

Proof of Theorem 3.3.1. Let P_{e} denote the probability that there exists some $\mathbf{x}, \mathbf{y} \in \mathbb{B}_R^n$ with $\|\mathbf{x}-\mathbf{y}\| > \epsilon$ and $\mathscr{A}(\mathbf{x}) = \mathscr{A}(\mathbf{y})$. Our goal is to show that $P_{\text{e}} \leq \eta$. Towards this end, let U be a δ-covering set for $\mathbb{B}_R^n$ with $|U| \leq (3R/\delta)^n$. By construction, for any $\mathbf{x}, \mathbf{y} \in \mathbb{B}_R^n$, there exist some $\mathbf{w}, \mathbf{z} \in U$ satisfying $\|\mathbf{x}-\mathbf{w}\| \leq \delta$ and $\|\mathbf{y}-\mathbf{z}\| \leq \delta$. In this case, if $\|\mathbf{x}-\mathbf{y}\| > \epsilon$ then

$$\|\mathbf{w}-\mathbf{z}\| \geq \|\mathbf{x}-\mathbf{y}\| - 2\delta > \epsilon - 2\delta.$$

Our goal is to upper bound the probability that there exists some $\mathbf{w}, \mathbf{z} \in U$ with $\|\mathbf{w}-\mathbf{z}\| \geq$

$\epsilon_0 = \epsilon - 2\delta$ and $\mathscr{A}(\mathbf{u}) = \mathscr{A}(\mathbf{v})$ for some $\mathbf{u} \in B_\delta(\mathbf{w})$ and $\mathbf{v} \in B_\delta(\mathbf{z})$. Said differently, we would like to bound the probability that there exists a $\mathbf{w}, \mathbf{z} \in U$ with $\|\mathbf{w}-\mathbf{z}\| \geq \epsilon_0$ for which $B_\delta(\mathbf{w})$ and $B_\delta(\mathbf{z})$ are not separated by any of the m hyperplanes.

Let $P_m(\mathbf{w}, \mathbf{z})$ denote the probability that $B_\delta(\mathbf{w})$ and $B_\delta(\mathbf{z})$ are not separated by any of the m hyperplanes for a fixed $\mathbf{w}, \mathbf{z} \in U$ with $\|\mathbf{w} - \mathbf{z}\| \geq \epsilon_0$. Lemma 3.3.2 controls this probability for a single hyperplane, yielding a bound of

$$1 - P_{\text{sep}} \leq 1 - \frac{\epsilon_0 - \delta\sqrt{2n}}{22\sqrt{\pi}e^{5/2}R}.$$

Since the $(\mathbf{p}_i, \mathbf{q}_i)$ are independent, we obtain

$$P_m(\mathbf{w}, \mathbf{z}) \leq \left(1 - \frac{\epsilon_0 - \delta\sqrt{2n}}{22\sqrt{\pi}e^{5/2}R}\right)^m. \tag{3.8}$$

Since we are interested in the event that there exists *any* $\mathbf{w}, \mathbf{z} \in U$ with $\|\mathbf{w} - \mathbf{z}\| \geq \epsilon_0$ for which $B_\delta(\mathbf{w})$ and $B_\delta(\mathbf{z})$ are separated by *none* of the m hyperplanes, we use the fact that there are at most $(3R/\delta)^{2n}$ such pairs $\mathbf{w}, \mathbf{z}$ and combine a union bound with (3.8) to obtain

$$P_{\text{e}} \leq \left(\frac{3R}{\delta}\right)^{2n}\left(1 - \frac{\epsilon_0 - \delta\sqrt{2n}}{22\sqrt{\pi}e^{5/2}R}\right)^m \leq \exp\left(2n\log\frac{3R}{\delta} - \frac{\left(\epsilon_0 - \delta\sqrt{2n}\right)m}{22\sqrt{\pi}e^{5/2}R}\right), \tag{3.9}$$

which follows from $(1 - x) \leq e^{-x}$. Bounding the right-hand side of (3.9) by η, we obtain

$$2n\log\frac{3R}{\delta} - \frac{\left(\epsilon_0 - \delta\sqrt{2n}\right)m}{22\sqrt{\pi}e^{5/2}R} \leq \log\eta. \tag{3.10}$$

If we now make the substitutions $\epsilon_0 = \epsilon - 2\delta$ and $\delta = \epsilon/(4 + \sqrt{8n})$, then we have that $\epsilon_0 - \delta\sqrt{n} = \epsilon/2$ and thus we can reduce (3.10) to

$$2n\log\frac{3R(4 + \sqrt{8n})}{\epsilon} - \frac{\epsilon m}{44\sqrt{\pi}e^{5/2}R} \leq \log\eta.$$

By rearranging, we see that this is equivalent to

$$m \geq 44\sqrt{\pi}e^{5/2}\frac{R}{\epsilon}\left(2n\log\frac{3R(4 + \sqrt{8n})}{\epsilon} + \log\frac{1}{\eta}\right). \tag{3.11}$$

One can easily show that (3.4) implies (3.11) for an appropriate choice of C. □

We now show that the result in Theorem 3.3.1 is optimal in the sense that *any* set of comparisons which can guarantee a uniform recovery of all $\mathbf{x} \in \mathbb{B}_R^n$ to accuracy ϵ will require a number of comparisons on the same order as that required in Theorem 3.3.1 (up to log factors).

Theorem 3.3.3. *For any configuration of m (inhomogeneous) hyperplanes in $\mathbb{R}^n$ dividing $\mathbb{B}_R^n$ into cells, if $m < \frac{2}{e}\frac{R}{\epsilon}n$, then there exist two points $\mathbf{x}, \mathbf{y} \in \mathbb{B}_R^n$ in the same cell such that $\|\mathbf{x} - \mathbf{y}\| \geq \epsilon$.*

Proof. We will use two facts. First, the number of cells (both bounded and unbounded) defined by m hyperplanes in $\mathbb{R}^n$ in general position[2] is given by

$$F_n(m) = \sum_{i=0}^{n} \binom{m}{i} \leq \left(\frac{em}{n}\right)^n < \left(\frac{2R}{\epsilon}\right)^n, \tag{3.12}$$

where the second inequality follows from the assumption that $m < 2Rn/e\epsilon$.

Second, for any convex set K we have the isodiametric inequality [144]: where $\mathrm{Diam}(K) = \sup_{x,y\in K}\|x - y\|$,

$$\left(\frac{\mathrm{Diam}(K)}{2}\right)^n \frac{\pi^{n/2}}{\Gamma(n/2+1)} \geq \mathrm{Vol}(K), \tag{3.13}$$

with equality when K is a ball. Since the entire volume of $\mathbb{B}_R^n$, denoted $\mathrm{Vol}(\mathbb{B}_R^n)$, is filled by at most $F_n(m)$ non-overlapping cells, there must exist at least one such cell K_0 with

$$\mathrm{Vol}(K_0) \geq \frac{\mathrm{Vol}(\mathbb{B}_R^n)}{F_n(m)} = \frac{\pi^{n/2}}{\Gamma(n/2+1)}\frac{R^n}{F_n(m)}. \tag{3.14}$$

Combining (3.13) with (3.14), we obtain

$$\left(\frac{\mathrm{Diam}(K_0)}{2}\right)^n \geq \frac{R^n}{F_n(m)},$$

which, together with (3.12), implies that

$$\mathrm{Diam}(K_0) \geq \frac{2R}{\sqrt[n]{F_n(m)}} > \epsilon.$$

[2]For non-general position, this is an upper bound [143].

Thus there are vectors $\mathbf{x}, \mathbf{y} \in K_0$ such that $\|\mathbf{x} - \mathbf{y}\| > \epsilon$. □

3.4 Stability in noise

So far, we have only considered the noise-free case. In most practical applications, observations may be corrupted by noise. We consider two scenarios; in the first, Gaussian noise is added prior to the sign(·) function in (3.3); in the second we make no assumption on the source of the errors and instead show the paired comparison observations are stable with respect to Euclidean distance. That is, two signals that have similar sign patterns are also nearby (and vice-versa). One can view this as a strengthening of the result in Theorem 3.3.1.

Throughout the following, we denote by d_H the Hamming distance, i.e., d_H counts the fraction of comparisons which differ between two sets of observations, here denoted $\mathscr{A}(\mathbf{x})$ and $\mathscr{A}(\mathbf{y})$:

$$d_H(\mathscr{A}(\mathbf{x}), \mathscr{A}(\mathbf{y})) := \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} |\mathscr{A}_i(\mathbf{x}) - \mathscr{A}_i(\mathbf{y})|. \tag{3.15}$$

3.4.1 Gaussian noise

Here we aim to understand how the paired comparisons change with the introduction of "pre-quantization" Gaussian noise. This will have the effect of causing some comparisons to be erroneous, where the probability of an error will be largest when $\mathbf{x}$ is equidistant from $\mathbf{p}_i$ and $\mathbf{q}_i$ and will decay as $\mathbf{x}$ moves away from this boundary.

Towards this end, recall that the observation model in (3.1) can be reduced to the form

$$\mathscr{A}_i(\mathbf{x}) = \text{sign}(q_i) \qquad q_i := \mathbf{a}_i^T \mathbf{x} - \tau_i. \tag{3.16}$$

In the noisy case, we will consider the observations

$$\bar{\mathscr{A}}_i(\mathbf{x}) = \text{sign}(\bar{q}_i) \qquad \bar{q}_i := \mathbf{a}_i^T \mathbf{x} - \tau_i + z_i = \bar{q}_i + z_i, \tag{3.17}$$

where $z_i \sim \mathcal{N}(0, \sigma_z^2)$. Note that since $\|\mathbf{a}_i\| = 1$, this model is equivalent to adding multivariate Gaussian noise directly to $\mathbf{x}$ with covariance $\sigma_z^2 I$. For a fixed $\mathbf{x}$, we can then quantify the probability that $d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x}))$ is large via the following theorem.

Theorem 3.4.1. *Suppose*[3] $n \geq 4$ *and fix* $\mathbf{x} \in \mathbb{B}_R^n$. *Let* $\mathscr{A}(\mathbf{x})$ *and* $\bar{\mathscr{A}}(\mathbf{x})$ *denote the collection of* m *observations defined as in* (3.16) *and* (3.17) *respectively, where the* $\{(\mathbf{p}_i, \mathbf{q}_i)\}_{i=1}^m$ *(and hence the* $\{(\mathbf{a}_i, \tau_i)\}_{i=1}^m$*) are generated as in Theorem 3.3.1. Then,*

$$\mathbb{E}\, d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x})) \leq \kappa_n(\sigma_z^2) \tag{3.18}$$

and

$$\mathbb{P}\left[d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x})) \geq \kappa_n(\sigma_z^2) + \zeta\right] \leq \exp(-2m\zeta^2), \tag{3.19}$$

where

$$\kappa_n(\sigma_z^2) := \sqrt{\frac{\sigma_z^2}{\sigma_z^2 + 2R^2/n + 4\|\mathbf{x}\|^2/n}}. \tag{3.20}$$

Proof. By Lemma 3.4.2, we have that $\mathbb{P}[\mathscr{A}_i(\mathbf{x}) \neq \bar{\mathscr{A}}_i(\mathbf{x})]$ is bounded by $\kappa_n(\sigma_z^2)$. Since the comparisons are independent, the expected number of sign mismatches is just the probability of a sign flip just computed, which establishes (3.18). The tail bound in (3.19) is a simple consequence of Hoeffding's inequality. □

To place this result in context, recall that $\tau_i \sim \mathcal{N}(0, R^2/n)$. Suppose that $\sigma_z^2 = c_0 R^2/n$. In this case one can bound (3.20) as

$$\sqrt{\frac{c_0}{c_0 + 6}} \leq \kappa_n(\sigma_z^2) \leq \sqrt{\frac{c_0}{c_0 + 2}}.$$

Intuitively, if c_0 is close to 1, then we would expect to lose a significant amount of information about $\mathbf{x}$, in which case $d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x}))$ could potentially be quite large. Indeed, if $c_0 > \frac{1}{2}$, then the lower bound above yields $\kappa_n(\sigma_z^2) > \frac{1}{2}$, meaning that our bound is essentially vacuous. In contrast, by letting c_0 grow small we can bound $\kappa_n(\sigma_z^2) \leq \sqrt{c_0/2}$ arbitrarily close to zero.

Lemma 3.4.2. *Suppose* $n \geq 4$. *Then* $\mathbb{P}[\mathscr{A}_i(\mathbf{x}) \neq \bar{\mathscr{A}}_i(\mathbf{x})] \leq \kappa_n(\sigma_z^2)$ *where* κ_n *is defined in* (3.20).

[3] For clarity, we focus on the $n \geq 4$ case. We consider the $n = 2$ and $n = 3$ cases separately because when $n \geq 4$ the probability distribution function of $\mathbf{a}_i^T \mathbf{x}$ is well-approximated by a Gaussian function but not for $n < 4$. We give alternative expressions for κ_n when $n = 2$ and $n = 3$ in Appendix 3.9.

Proof. The probability of a sign flip is given by

$$\mathbb{P}\left[q_i\bar{q}_i < 0\right] = \mathbb{P}\left[q_i < 0 \text{ and } \bar{q}_i > 0\right] + \mathbb{P}\left[q_i > 0 \text{ and } \bar{q}_i < 0\right].$$

Note that if we set $d_i = \mathbf{a}_i^T\mathbf{x}/\|\mathbf{x}\| \in [-1, 1]$, then we can write $q_i = d_i\|\mathbf{x}\| - \tau_i$ and $\bar{q}_i = d_i\|\mathbf{x}\| - \tau_i + z_i$. Thus, if $f_d(d_i)$, $f_\tau(\tau_i)$, and $f_z(z_i)$ denote the probability density functions for d_i, τ_i, and z_i, then since these random variables are independent we can write

$$\begin{aligned}
\mathbb{P}\left[q_i < 0 \text{ and } \bar{q}_i > 0\right] &= \mathbb{P}\left[d_i\|\mathbf{x}\| - \tau_i < 0 \text{ and } d_i\|\mathbf{x}\| - \tau_i + z_i > 0\right] \\
&= \int_{-1}^{1}\int_{d_i\|\mathbf{x}\|}^{\infty}\int_{-\infty}^{d_i\|\mathbf{x}\|-\tau_i} f_d(d_i)f_\tau(\tau_i)f_z(z_i)\,\mathrm{d}z_i\,\mathrm{d}\tau_i\,\mathrm{d}d_i \\
&= \int_{-1}^{1}\int_{d_i\|\mathbf{x}\|}^{\infty} f_d(d_i)f_\tau(\tau_i)\,\mathbb{P}\left[z_i > \tau_i - d_i\|\mathbf{x}\|\right]\mathrm{d}\tau_i\,\mathrm{d}d_i \\
&= \int_{-1}^{1}\int_{d_i\|\mathbf{x}\|}^{\infty} f_d(d_i)f_\tau(\tau_i)Q\left(\frac{\tau_i - d_i\|\mathbf{x}\|}{\sigma_z}\right)\mathrm{d}\tau_i\,\mathrm{d}d_i,
\end{aligned}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-x^2/2)\,\mathrm{d}x$, i.e., the tail probability for the standard normal distribution. Via a similar argument we have

$$\begin{aligned}
\mathbb{P}\left[q_i > 0 \text{ and } \bar{q}_i < 0\right] &= \mathbb{P}\left[d_i\|\mathbf{x}\| - \tau_i > 0 \text{ and } d_i\|\mathbf{x}\| - \tau_i + z_i < 0\right] \\
&= \int_{-1}^{1}\int_{-\infty}^{d_i\|\mathbf{x}\|}\int_{d_i\|\mathbf{x}\|-\tau_i}^{\infty} f_d(d_i)f_\tau(\tau_i)f_z(z_i)\,\mathrm{d}z_i\,\mathrm{d}\tau_i\,\mathrm{d}d_i \\
&= \int_{-1}^{1}\int_{-\infty}^{d_i\|\mathbf{x}\|} f_d(d_i)f_\tau(\tau_i)\,\mathbb{P}\left[z_i < \tau_i - d_i\|\mathbf{x}\|)\right]\mathrm{d}\tau_i\,\mathrm{d}d_i \\
&= \int_{-1}^{1}\int_{-\infty}^{d_i\|\mathbf{x}\|} f_d(d_i)f_\tau(\tau_i)Q\left(\frac{d_i\|\mathbf{x}\| - \tau_i}{\sigma_z}\right)\mathrm{d}\tau_i\,\mathrm{d}d_i.
\end{aligned}$$

Combining these we obtain

$$\begin{aligned}
\mathbb{P}[q_i\bar{q}_i < 0] &= \mathbb{P}\left[q_i < 0 \text{ and } \bar{q}_i > 0\right] + \mathbb{P}\left[q_i > 0 \text{ and } \bar{q}_i < 0\right] \\
&= \int_{-1}^{1}\int_{-\infty}^{\infty} f_d(d_i)f_\tau(\tau_i)Q\left(\frac{|d_i\|\mathbf{x}\| - \tau_i|}{\sigma_z}\right)\mathrm{d}\tau_i\,\mathrm{d}d_i \\
&= 2\int_{0}^{1}\int_{-\infty}^{\infty} f_d(d_i)f_\tau(\tau_i)Q\left(\frac{|d_i\|\mathbf{x}\| - \tau_i|}{\sigma_z}\right)\mathrm{d}\tau_i\,\mathrm{d}d_i,
\end{aligned}$$

following from the symmetry of $f_d(\cdot)$. Using the bound $Q(x) \le \frac{1}{2}\exp(-x^2/2)$ (see (13.48)

of [145]), and recalling that $\tau_i \sim \mathcal{N}(0, 2R^2/n)$, we have that

$$\mathbb{P}[q_i \bar{q}_i < 0] \leq \frac{1}{R}\sqrt{\frac{n}{\pi}} \int_0^1 \int_{-\infty}^{\infty} f_d(d_i) \exp\left(-\frac{(d_i \|\mathbf{x}\| - \tau_i)^2}{2\sigma_z^2} - \frac{n\tau_i^2}{4R^2} \right) \mathrm{d}\tau_i \, \mathrm{d}d_i. \quad (3.21)$$

The remainder of the proof (given in Section 3.9) is obtained by bounding this integral. Note that in general, we have $\frac{1}{2}(d_i+1) \sim \mathrm{Beta}((n-1)/2, (n-1)/2)$, but d_i is asymptotically normal with variance $1/n$ [146]. For $n \geq 4$, we use the simple upper bound

$$\begin{aligned} f_d(d_i) &= \left[B\left(\frac{n-1}{2}, \frac{n-1}{2} \right) \right]^{-1} \left(\frac{1+d_i}{2} \frac{1-d_i}{2} \right)^{(n-3)/2} \\ &\leq \left[\frac{\sqrt{2\pi} \frac{n-1}{2}^{(n-2)/2} \frac{n-1}{2}^{(n-2)/2}}{(n-1)^{n-1-1/2}} \right]^{-1} \left(\frac{1-d_i^2}{4} \right)^{(n-3)/2} \\ &= \left[\frac{\sqrt{2\pi}}{2^{n-2}\sqrt{n-1}} \right]^{-1} \frac{1}{2^{n-3}} \exp(-(n-3)d_i^2/2) \\ &= \frac{\sqrt{n-1}}{2\sqrt{2\pi}} \exp(-(n-3)d_i^2/2) \\ &\leq \frac{\sqrt{n}}{2\sqrt{2\pi}} \exp(-nd_i^2/8). \end{aligned} \quad (3.22)$$

This follows from the standard inequalities $B(x, y) \geq \sqrt{2\pi} x^{x-1/2} y^{y-1/2}/(x+y)^{x+y-1/2}$ [e.g., 147] and $1 - x \leq \exp(-x)$. □

3.4.2 Stable embedding

Here we show that given enough comparisons there is an approximate embedding of the preference space into $\{-1, 1\}^m$ via our model. Theorem 3.4.3 states that if $\mathbf{x}$ and $\mathbf{y}$ are sufficiently close, then the respective comparison patterns $\mathcal{A}(\mathbf{x})$ and $\mathcal{A}(\mathbf{y})$ closely align. In contrast with Theorem 3.4.1, Theorem 3.4.3 is a purely geometric statement which makes no assumptions on any particular noise model. Note also that Theorem 3.4.3 applies uniformly *for all* $\mathbf{x}$ and $\mathbf{y}$.

Theorem 3.4.3. *Let $\eta, \zeta > 0$ be given. Let $\mathcal{A}(\mathbf{x})$ denote the collection of m observations*

defined as in Theorem 3.3.1. There exist constants C_1, c_1, C_2, c_2 such that if

$$m \geq \frac{1}{2\zeta^2}\left(2n \log \frac{3\sqrt{n}}{\zeta} + \log \frac{2}{\eta}\right), \tag{3.23}$$

then with probability at least $1 - \eta$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{B}_R^n$ we have

$$C_1 \frac{\|\mathbf{x} - \mathbf{y}\|}{R} - c_1\zeta \leq d_H(\mathscr{A}(\mathbf{x}), \mathscr{A}(\mathbf{y})) \leq C_2 \frac{\|\mathbf{x} - \mathbf{y}\|}{R} + c_2\zeta. \tag{3.24}$$

This result implies that the fraction of differences in the set of observed comparisons between $\mathbf{x}$ and $\mathbf{y}$ will be constrained to within a constant factor of the Euclidean distance, plus an additive error approximately proportional to $1/\sqrt{m}$. At first glance, this seems worse than the result of Theorem 3.3.1, which suggests the rate $1/m$. However, Theorem 3.4.3 comes with much greater flexibility in that Theorem 3.3.1 only concerns the case where $d_H(\mathscr{A}(\mathbf{x}), \mathscr{A}(\mathbf{y})) = 0$. Like Theorem 3.3.1, this result applies *for all* $\mathbf{x}$ on the same randomly drawn set of items.

In the context of a hypothetical recovery problem, suppose $\mathbf{x}$ is a parameter of interest and $\mathbf{y}$ is an estimate produced by any algorithm. Then (3.24) says that if we want to recover $\mathbf{x}$ to within error ϵ, the algorithm should look for vectors $\mathbf{y}$ which is have up to $O(\epsilon)$ incorrect comparisons. Likewise, if a $\mathbf{y}$ can be found having up to $O(\epsilon)$ comparison errors, we have the same $O(\epsilon)$ guarantee on the Euclidean error of the estimate.

It is also instructive to consider this result next to Theorem 3.4.1 which also predicts the fraction of sign mismatches generated by noise up to an additive constant which is proportional to $1/\sqrt{m}$. If in a particular application the noise is expected to be Gaussian, the bound (3.19) can be used as guidance when using (3.24) since together they predict the fraction of comparison errors which is unavoidable. In this case, Theorem 3.3.1 would be inappropriate because it may be impossible to find a $\mathbf{y}$ such that $d_H(\bar{\mathscr{A}}(\mathbf{x}), \mathscr{A}(\mathbf{y})) = 0$.

Proof. By Lemma 3.4.4, for any fixed pair $\mathbf{w}, \mathbf{z} \in \mathbb{B}_R^n$ we have bounds on the Hamming distance that hold with probability at least $1 - 2\exp(-2\zeta^2 m)$, for all $\mathbf{u} \in B_\delta(\mathbf{w})$ and $\mathbf{v} \in B_\delta(\mathbf{z})$. Recall that the radius R ball can be covered with a set U of radius δ balls with $|U| \leq (3R/\delta)^n$.

Thus, by a union bound we have that with probability at least $1 - 2(3R/\delta)^{2n}\exp(-2\zeta^2 m)$, for *any* $\mathbf{w}, \mathbf{z} \in U$,

$$\frac{1}{22e^{5/2}\sqrt{\pi}}\left(\frac{\|\mathbf{w}-\mathbf{z}\|}{R} - \frac{\delta\sqrt{2n}}{R}\right) - \zeta \leq d_H(\mathscr{A}(\mathbf{u}), \mathscr{A}(\mathbf{v})) \leq \sqrt{\frac{2}{\pi}}\left(\frac{\|\mathbf{w}-\mathbf{z}\|}{R} + \frac{\delta\sqrt{n}}{R}\right) + \zeta,$$

for all $\mathbf{u} \in B_\delta(\mathbf{w})$ and $\mathbf{v} \in B_\delta(\mathbf{z})$. Since $\|\mathbf{x}-\mathbf{y}\| - 2\delta \leq \|\mathbf{w}-\mathbf{z}\| \leq \|\mathbf{x}-\mathbf{y}\| + 2\delta$, this implies that

$$\frac{1}{22e^{5/2}\sqrt{\pi}}\left(\frac{\|\mathbf{x}-\mathbf{y}\| - 2\delta}{R} - \frac{\delta\sqrt{2n}}{R}\right) - \zeta \leq d_H(\mathscr{A}(\mathbf{x}), \mathscr{A}(\mathbf{y})) \leq \sqrt{\frac{2}{\pi}}\left(\frac{\|\mathbf{x}-\mathbf{y}\| + 2\delta}{R} + \frac{\delta\sqrt{n}}{R}\right) + \zeta,$$

Letting $\delta = \zeta R/\sqrt{n}$ and setting C_1, c_1, C_2, c_1 appropriately[4] this reduces to (3.24). Lower bounding the probability by $1 - \eta$, we obtain

$$2(3\sqrt{n}/\zeta)^{2n}\exp(-2\zeta^2 m) \leq \eta.$$

Rearranging yields (3.23). □

Lemma 3.4.4. *Let $\mathbf{w}, \mathbf{z} \in \mathbb{B}_R^n$ be distinct and fixed, and let $\delta, \zeta > 0$ be given. Let $\mathscr{A}(\mathbf{x})$ denote the collection of m observations defined as in Theorem 3.3.1, and let $B_\delta(\cdot)$ be defined as in Lemma 3.3.2. Then for all $\mathbf{u} \in B_\delta(\mathbf{w})$ and $\mathbf{v} \in B_\delta(\mathbf{z})$,*

$$\frac{1}{22e^{5/2}\sqrt{\pi}}\left(\frac{\|\mathbf{w}-\mathbf{z}\|}{R} - \frac{\delta\sqrt{2n}}{R}\right) - \zeta \leq d_H(\mathscr{A}(\mathbf{u}), \mathscr{A}(\mathbf{v})) \leq \sqrt{\frac{2}{\pi}}\left(\frac{\|\mathbf{w}-\mathbf{z}\|}{R} + \frac{\delta\sqrt{n}}{R}\right) + \zeta,$$

with probability at least $1 - \exp(-2\zeta^2 m)$.

Proof. Fix $\delta > 0$ and let $\mathbf{u} \in B_\delta(\mathbf{w}), \mathbf{v} \in B_\delta(\mathbf{z})$. Recall that the Hamming distance d_H is a sum of independent and identically distributed Bernoulli random variables and we may bound it using Hoeffding's inequality. Since our probabilistic upper and lower bounds must hold for all $\mathbf{u}, \mathbf{v}$ as described above, we introduce quantities L_0 and L_1 which represent two

[4] We set $C_1 = 1/22e^{5/2}\sqrt{\pi}$ and $C_2 = \sqrt{2/\pi}$. We may set $c_1 = 1 + 1/11e^{5/2}\sqrt{\pi} + \sqrt{2/\pi}$ and $c_2 = 1 + 3\sqrt{2/\pi}$ to obtain constants that are valid for all n – improved values are possible for large n.

"extreme cases" of the Bernoulli variables:

$$L_0 := \sup_{\mathbf{u}\in B_\delta(\mathbf{w}),\mathbf{v}\in B_\delta(\mathbf{z})} \frac{1}{2m}\sum_{i=1}^{m} |\mathscr{A}_i(\mathbf{u}) - \mathscr{A}_i(\mathbf{v})|$$

$$L_1 := \inf_{\mathbf{u}\in B_\delta(\mathbf{w}),\mathbf{v}\in B_\delta(\mathbf{z})} \frac{1}{2m}\sum_{i=1}^{m} |\mathscr{A}_i(\mathbf{u}) - \mathscr{A}_i(\mathbf{v})|.$$

Then we have

$$L_1 \leq d_H(\mathscr{A}(\mathbf{u}), \mathscr{A}(\mathbf{v})) \leq L_0$$

Denote $P_0 = 1 - \mathbb{E}\, L_0$ and $P_1 = \mathbb{E}\, L_1$, i.e.,

$$P_0 = \mathbb{P}\left[\forall \mathbf{u} \in B_\delta(\mathbf{w}), \forall \mathbf{v} \in B_\delta(\mathbf{z}) : \mathscr{A}_i(\mathbf{u}) = \mathscr{A}_i(\mathbf{v})\right]$$

$$P_1 = \mathbb{P}\left[\forall \mathbf{u} \in B_\delta(\mathbf{w}), \forall \mathbf{v} \in B_\delta(\mathbf{z}) : \mathscr{A}_i(\mathbf{u}) \neq \mathscr{A}_i(\mathbf{v})\right].$$

By Hoeffding's inequality,

$$\mathbb{P}\left[L_0 > (1 - P_0) + \zeta\right] \leq \exp(-2m\zeta^2)$$

$$\mathbb{P}\left[L_1 < P_1 - \zeta\right] \leq \exp(-2m\zeta^2).$$

Hence, with probability at least $1 - 2\exp(-2m\zeta^2)$,

$$P_1 - \zeta \leq d_H(\mathscr{A}(\mathbf{u}), \mathscr{A}(\mathbf{v})) \leq (1 - P_0) + \zeta.$$

The result follows directly from this combined with the facts that from Lemma 3.3.2 we have

$$P_1 \geq \frac{1}{22e^{5/2}\sqrt{\pi}}\left(\frac{\|\mathbf{w} - \mathbf{z}\|}{R} - \frac{\delta\sqrt{2n}}{R}\right),$$

and from Lemma 3.4.5 we have

$$1 - P_0 \leq \sqrt{\frac{2}{\pi}}\left(\frac{\|\mathbf{w} - \mathbf{z}\|}{R} + \frac{\delta\sqrt{n}}{R}\right).$$

□

Lemma 3.4.5. *Let $\mathbf{w}, \mathbf{z} \in \mathbb{B}_R^n$ be distinct and fixed, and let $\delta > 0$ be given. Let $\mathscr{A}(\mathbf{x})$ denote the collection of m observations defined as in Theorem 3.3.1, and let $B_\delta(\cdot)$ be defined as*

in Lemma 3.3.2. Denote by P_0 the probability that $B_\delta(\mathbf{w})$ and $B_\delta(\mathbf{z})$ are not separated by hyperplane i, i.e.,

$$P_0 = \mathbb{P}\left[\forall \mathbf{u} \in B_\delta(\mathbf{w}), \forall \mathbf{v} \in B_\delta(\mathbf{z}) : \mathscr{A}_i(\mathbf{u}) = \mathscr{A}_i(\mathbf{v})\right].$$

Then

$$1 - P_0 \leq \sqrt{\frac{2}{\pi}}\left(\frac{\|\mathbf{w}-\mathbf{z}\|}{R} + \frac{\delta\sqrt{n}}{R}\right).$$

Proof. We need an upper bound on

$$1 - P_0 = \mathbb{P}\left[\mathscr{A}_i(\mathbf{u}) \neq \mathscr{A}_i(\mathbf{v}) \text{ for some } \mathbf{u} \in B_\delta(\mathbf{w}), \mathbf{v} \in B_\delta(\mathbf{z})\right].$$

Suppose for now that $\mathbf{a}$ is fixed and without loss of generality that $\mathbf{a}^T\mathbf{w} > \mathbf{a}^T\mathbf{z}$. Then this probability is simply

$$\begin{aligned}\mathbb{P}\left[\mathbf{a}^T\mathbf{v} < \tau < \mathbf{a}^T\mathbf{u} \text{ for some } \mathbf{u} \in B_\delta(\mathbf{w}), \mathbf{v} \in B_\delta(\mathbf{z})\right] &= \mathbb{P}\left[\min_{\mathbf{v}\in B_\delta(\mathbf{z})} \mathbf{a}^T\mathbf{v} < \tau < \max_{\mathbf{u}\in B_\delta(\mathbf{w})} \mathbf{a}^T\mathbf{u}\right] \\ &\leq \mathbb{P}\left[\mathbf{a}^T\mathbf{z} - \delta < \tau < \mathbf{a}^T\mathbf{w} + \delta\right],\end{aligned}$$

since by Cauchy-Schwarz we have

$$\min_{\mathbf{v}\in B_\delta(\mathbf{z})} \mathbf{a}^T\mathbf{v} \geq \mathbf{a}^T\mathbf{z} - \delta \qquad \text{and} \qquad \max_{\mathbf{u}\in B_\delta(\mathbf{w})} \mathbf{a}^T\mathbf{u} \leq \mathbf{a}^T\mathbf{w} + \delta.$$

Thus, recalling that $\tau_i \sim \mathcal{N}(0, R^2/n)$, from Lemma 3.8.1 we have

$$\begin{aligned}\mathbb{P}\left[\mathbf{a}^T\mathbf{z} - \delta < \tau < \mathbf{a}^T\mathbf{w} + \delta\right] &= \Phi\left(\frac{\mathbf{a}^T\mathbf{w} + \delta}{R/\sqrt{n}}\right) - \Phi\left(\frac{\mathbf{a}^T\mathbf{z} - \delta}{R/\sqrt{n}}\right) \\ &\leq \frac{1}{R}\sqrt{\frac{n}{2\pi}}\left(\mathbf{a}^T(\mathbf{w}-\mathbf{z}) + 2\delta\right).\end{aligned}$$

Similarly, for $\mathbf{a}^T\mathbf{w} < \mathbf{a}^T\mathbf{z}$ we have

$$\mathbb{P}\left[\mathbf{a}^T\mathbf{w} - \delta < \tau < \mathbf{a}^T\mathbf{z} + \delta\right] \leq \frac{1}{R}\sqrt{\frac{n}{2\pi}}\left(\mathbf{a}^T(\mathbf{z}-\mathbf{w}) + 2\delta\right).$$

Combining these we have

$$\begin{aligned}1 - P_0 &\leq \int_{\mathbb{S}^{n-1}} \frac{1}{R}\sqrt{\frac{n}{2\pi}} \left(|\mathbf{a}^T(\mathbf{w}-\mathbf{z})| + 2\delta\right) \nu(\mathrm{d}\mathbf{a}) \\ &= \frac{1}{R}\sqrt{\frac{n}{2\pi}} \int_{\mathbb{S}^{n-1}} |\mathbf{a}^T(\mathbf{w}-\mathbf{z})| \, \nu(\mathrm{d}\mathbf{a}) + \frac{2\delta}{R}\sqrt{\frac{n}{2\pi}} \\ &= \frac{\sqrt{2n}}{R\pi} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})} \|\mathbf{w}-\mathbf{z}\| + \frac{\delta}{R}\sqrt{\frac{2n}{\pi}},\end{aligned}$$

where the last equality is proven in Lemma 3.8.2. The lemma then follows from the facts that $\frac{\Gamma(1/2)}{\Gamma(1)} = \sqrt{\pi}$ and $\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})} \leq \frac{2}{\sqrt{2n-1}} \leq \sqrt{\frac{\pi}{n}}$ for $n \geq 2$ [148, (2.20)]. □

3.5 Estimation guarantees

3.5.1 Estimation algorithms

In the noise-free setting, given a set of comparisons $\mathscr{A}(\mathbf{x})$, we may produce an estimate $\hat{\mathbf{x}}$ by finding *any* $\hat{\mathbf{x}} \in \mathbb{B}_R^n$ satisfying $\mathscr{A}(\hat{\mathbf{x}}) = \mathscr{A}(\mathbf{x})$. A simple approach is the following convex program:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\|^2 \quad \text{subject to} \quad \mathscr{A}_i(\mathbf{x})(\mathbf{a}_i^T\mathbf{w} - \tau_i) \geq 0 \quad \forall i \in [m]. \tag{3.25}$$

This is relatively easy to solve since the constraints are simple linear inequalities and the feasible region is convex. Note that (3.25) is guaranteed to satisfy $\hat{\mathbf{x}} \in \mathbb{B}_R^n$ since $\mathbf{x} \in \mathbb{B}_R^n$ and $\mathbf{x}$ is feasible, so that $\|\hat{\mathbf{x}}\| \leq \|\mathbf{x}\| \leq R$. In this case we may apply Theorem 3.3.1 to argue that if m obeys the bound in (3.4), then $\|\hat{\mathbf{x}} - \mathbf{x}\| \leq \epsilon$.

However, in most practical applications, observations are likely to be corrupted by noise leading to inconsistencies. Any errors in the observations $\mathscr{A}(\mathbf{x})$ would make strictly enforcing $\mathscr{A}(\hat{\mathbf{x}}) = \mathscr{A}(\mathbf{x})$ a questionable goal since, among other drawbacks, $\mathbf{x}$ itself would become infeasible. In fact, in this case we cannot even necessarily guarantee that (3.25) has *any* feasible solutions. In the noisy case we instead use a relaxation inspired by the extended ν-SVM of [149], which introduces slack variables $\xi_i \geq 0$ and is controlled by the parameter ν. Specifically, we denote by $\bar{\mathscr{A}}(\mathbf{x})$ the collection of (potentially) corrupted measurements,

and we solve

$$\begin{aligned} \underset{\widehat{\mathbf{w}}\in\mathbb{R}^{n+1},\boldsymbol{\xi}\in\mathbb{R}^m,\rho\in\mathbb{R}}{\text{minimize}} \quad & -\nu\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & \bar{\mathscr{A}}_i(\mathbf{x})([\mathbf{a}_i^T, -\tau_i]\widehat{\mathbf{w}}) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \forall i \in [m], \\ & \|\widehat{\mathbf{w}}[1:n]\|^2 \leq \tfrac{2R^2}{1+R^2}, \quad \text{and} \quad \|\widehat{\mathbf{w}}\|^2 = 2. \end{aligned} \tag{3.26}$$

Finally, we set $\hat{\mathbf{x}} = \widehat{\mathbf{w}}[1,\ldots,n]/\widehat{\mathbf{w}}[n+1]$. The additional constraint $\|\widehat{\mathbf{w}}[1:n]\|^2 \leq \frac{2R^2}{1+R^2}$ ensures that $\|\hat{\mathbf{x}}\| \leq R$. Note that an important difference between the extended ν-SVM and (3.26) is that there is no "offset" parameter to be optimized over. That is, if we interpret $[\mathbf{a}_i, -\tau_i]$ as "training examples," then $\mathbf{w} := [\mathbf{x}, 1] \in \mathbb{R}^{n+1}$ corresponds to a *homogeneous* linear classifier. Note that in the absence of comparison errors, setting $\nu = 0$, we would have a feasible solution with $\xi_i = 0$.

Unfortunately, (3.26) is not convex and a unique global minimum cannot be guaranteed, i.e., there may be multiple solutions $\hat{\mathbf{x}}$. Nevertheless, the following result shows that any local minimum will have certain desirable properties, and in the process also provides guidance on choosing the parameter ν. Combined with our previous results, this also allows us to give recovery guarantees.

Proposition 3.5.1. *At* any *local minimum* $\hat{\mathbf{x}}$ *of* (3.26)*, we have* $\frac{1}{m}|\{i : \xi_i > 0\}| \leq \nu$. *If the corresponding* $\rho > 0$*, this further implies that* $d_H(\mathscr{A}(\hat{\mathbf{x}}), \bar{\mathscr{A}}(\mathbf{x})) \leq \nu$.

Proof. This proof follows similarly to that of Proposition 7.5 of [150], except applied to the extended ν-SVM of [149] and with the removal of the hyperplane bias term. Specifically, we first form the Lagrangian of (3.26):

$$\begin{aligned} L(\widehat{\mathbf{w}}, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \delta) = -\nu\rho + \frac{1}{m}\sum_i \xi_i - \sum_i (\alpha_i(\bar{\mathscr{A}}_i(\mathbf{x})[\mathbf{a}_i, -\tau_i]^T\widehat{\mathbf{w}} - \rho + \xi_i) + \beta_i\xi_i) \\ + \gamma\Big(\frac{2R^2}{1+R^2} - \|\widehat{\mathbf{w}}[1:n]\|^2\Big) - \delta(1 - \|\widehat{\mathbf{w}}\|^2). \end{aligned}$$

We define the functions corresponding to the equality constraints (h_1) and inequality con-

straints (g_i) as follows:

$$h_1(\mathbf{w}, \boldsymbol{\xi}, \rho) := (1 - \|\mathbf{w}\|^2)$$

$$g_{i\in[2m+1]}(\mathbf{w}, \boldsymbol{\xi}, \rho) := \begin{cases} \bar{\mathscr{A}_i}(\mathbf{x})[\mathbf{a}_i, -\tau_i]^T \widehat{\mathbf{w}} - \rho + \xi_i & i \in [1, m] \\ \xi_i & i \in [m+1, 2m] \\ -\left(\frac{2R^2}{1+R^2} - \|\widehat{\mathbf{w}}[1 : n]\|^2\right) & i = 2m+1. \end{cases}$$

Consider the $n + m + 2$ variables $(\widehat{\mathbf{w}}, \boldsymbol{\xi}, \rho)$. The gradient corresponding to the equality constraint, $\boldsymbol{\nabla}\mathbf{h}_1$, involves only the first $n+1$ variables. Thus, there exists an $m+1$ dimensional subspace $\mathscr{D} \subset \mathbb{R}^{n+m+2}$ where for any $\mathbf{d} \in \mathscr{D}$, $\boldsymbol{\nabla}\mathbf{h}_1^T\mathbf{d} = 0$. The gradients corresponding to the $2m + 1$ inequality constraints are given in the $(2m + 1) \times (n + m + 2)$ matrix

$$\mathbf{G} := \left[\begin{array}{c} \boldsymbol{\nabla}\mathbf{g}_1^T \\ \vdots \\ \boldsymbol{\nabla}\mathbf{g}_m^T \\ \hline \boldsymbol{\nabla}\mathbf{g}_{m+1}^T \\ \vdots \\ \boldsymbol{\nabla}\mathbf{g}_{2m}^T \\ \hline \boldsymbol{\nabla}\mathbf{g}_{2m+1}^T \end{array}\right] = \left[\begin{array}{c|cccc} \cdots & -1 & \cdots & 0 & 1 \\ \cdots & \vdots & \ddots & \vdots & \vdots \\ \leftarrow (n+1) \rightarrow & 0 & \cdots & -1 & 1 \\ \hline \text{irrelevant} & -1 & \cdots & 0 & 0 \\ \cdots & \vdots & \ddots & \vdots & \vdots \\ \cdots & 0 & \cdots & -1 & 0 \\ \hline \leftarrow\ \widehat{\mathbf{w}}[1 : n]\ \rightarrow\ 0 & 0 & \cdots & 0 & 0 \end{array}\right].$$

Since there is a $\mathbf{d} \in \mathscr{D}$ such that $(\mathbf{Gd})[i] < 0$ for all i (for example, $\mathbf{d} = [0, \ldots, 0|1, \ldots, 1, -1]$), the Mangasarian–Fromovitz constraint qualifications hold and we have the following first-order necessary conditions for local minima [see e.g., 151],

$$\frac{\partial L}{\partial \rho} = -\nu + \sum \alpha_i \quad \Longrightarrow \quad \sum \alpha_i = \nu$$

and

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{m} - \alpha_i - \beta_i = 0 \quad \Longrightarrow \quad \alpha_i + \beta_i = \frac{1}{m}.$$

Since $\sum_{i=1}^m \alpha_i = \nu$, at most a fraction of ν can have $\alpha_i = 1/m$. Now, any i such that $\xi_i > 0$ must have $\alpha_i = 1/m$ since by complimentary slackness, $\beta_i = 0$. Hence, ν is an upper bound

on the fraction of ξ such that $\xi_i > 0$.

Finally, note that if $\rho > 0$, then $\xi_i = 0$ implies $\bar{\mathscr{A}_i}(\mathbf{x})([\mathbf{a}_i^T, -\tau_i]\widehat{\mathbf{w}}) \geq \rho - \xi_i > 0$. Hence, the fraction of ξ such that $\xi_i > 0$ is an upper bound for $d_H(\mathscr{A}(\hat{\mathbf{x}}), \bar{\mathscr{A}}(\mathbf{x}))$. □

3.5.2 Estimation guarantees

We now show how the results of Theorems 3.4.1 and 3.4.3 can be combined with Proposition 3.5.1 to give recovery guarantees on $\|\hat{\mathbf{x}} - \mathbf{x}\|$ when (3.26) is used for recovery under realistic noisy observation models. We consider three basic noise models. In the first, an arbitrary (but small) fraction of comparisons are reversed. We then consider the implication of this result in the context of two other noise models, one where Gaussian noise is added to either the underlying $\mathbf{x}$ or to the comparisons "pre-quantization," that is, directly to $(\mathbf{a}_i^T\mathbf{x} - \tau_i)$, and another where the observations are generated using an arbitrary (but bounded) perturbation of $\mathbf{x}$. We will ultimately see that largely similar guarantees are possible in all three cases.

In our analysis of all three settings, we will use the fact that from the lower bound of Theorem 3.4.3 we have

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{R} \leq \frac{d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x})) + c_1\zeta}{C_1} \tag{3.27}$$

with probability at least $1 - \eta$ provided that m is sufficiently large, e.g., by taking

$$m = \frac{1}{2\zeta^2}\left(2n\log\frac{3\sqrt{n}}{\zeta} + \log\frac{2}{\eta}\right). \tag{3.28}$$

Note that by setting $\beta = \frac{9n}{\zeta^2}\left(\frac{2}{\eta}\right)^{1/n}$, we can rearrange (3.28) to be of the form

$$18m\left(\frac{2}{\eta}\right)^{1/n} = \beta\log\beta,$$

which implies that

$$\beta = \frac{18m(2/\eta)^{1/n}}{W\left(18m(2/\eta)^{1/n}\right)},$$

where $W(\cdot)$ denotes the Lambert W function. Using the fact that $W(x) \leq \log(x)$ for $x \geq e$

and substituting back in for β, we have

$$\zeta \leq \sqrt{\frac{n\log(18m) + \log(2/\eta)}{2m}}$$

under the mild assumption that $m \geq \frac{e}{18}(\frac{\eta}{2})^{1/n}$. Substituting this in to (3.27) yields

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{R} \leq \frac{d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x}))}{C_1} + \frac{c_1}{C_1}\sqrt{\frac{n\log(18m) + \log(2/\eta)}{2m}}. \tag{3.29}$$

We use this bound repeatedly below.

Noise model 1..— In the first noise model, we suppose that an adversary is allowed to arbitrarily flip a fraction κ of measurements, where we assume κ is known (or can be bounded). This would seem to be a challenging setting, but in fact a guarantee under this model follows immediately from the lower bound in Theorem 3.4.3. Specifically, suppose that $\mathscr{A}(\mathbf{x})$ represents the noise-free comparisons, and we receive instead $\bar{\mathscr{A}}(\mathbf{x})$, where $d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x})) \leq \kappa$.

Consider using (3.26) to produce an $\hat{\mathbf{x}}$ setting $\nu = \kappa$. If $\hat{\mathbf{x}}$ is a local minimum for (3.26) with $\rho > 0$, Proposition 3.5.1 implies that $d_H(\mathscr{A}(\hat{\mathbf{x}}), \bar{\mathscr{A}}(\mathbf{x})) \leq \kappa$. Thus, by the triangle inequality,

$$d_H(\mathscr{A}(\hat{\mathbf{x}}), \mathscr{A}(\mathbf{x})) \leq d_H(\mathscr{A}(\hat{\mathbf{x}}), \bar{\mathscr{A}}(\mathbf{x})) + d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x})) \leq 2\kappa.$$

Plugging this into (3.29) we have that with probability at least $1 - \eta$

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{R} \leq \frac{2\kappa}{C_1} + \frac{c_1}{C_1}\sqrt{\frac{n\log(18m) + \log(2/\eta)}{2m}}. \tag{3.30}$$

We emphasize the power of this result—the adversary may flip not merely a random fraction of comparisons, but an *arbitrary* set of comparisons. Moreover, this holds uniformly for all $\mathbf{x}$ and $\hat{\mathbf{x}}$ simultaneously (with high probability).

Noise model 2..— Here we model errors as being generated by adding i.i.d. Gaussian before the sign$(\cdot)$ function, as described in Section 3.4.1, i.e.,

$$\bar{\mathscr{A}}_i(\mathbf{x}) = \text{sign}(\mathbf{a}_i^T\mathbf{x} - \tau_i + z_i),$$

where $z_i \sim \mathcal{N}(0, \sigma_z^2)$. Note that this model is equivalent to the Thurstone model of comparative judgment [152], and causes a predictable probability of error depending the geometry of the set of items. Specifically, comparisons which are "decisive," i.e., whose hyperplane lies far from $\mathbf{x}$, are unlikely to be affected by this noise. Conversely, comparisons which are nearly even are quite likely to be affected.

Under the random observation model considered in this chapter, by Theorem 3.4.1 we have that, with probability at least $1 - \eta$,

$$d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x})) \leq \kappa_n(\sigma_z^2) + \sqrt{\frac{\log(1/\eta)}{2m}},$$

where

$$\kappa_n(\sigma_z^2) = \sqrt{\frac{\sigma_z^2}{\sigma_z^2 + 2R^2/n + 4\|\mathbf{x}\|^2/n}} \leq \sqrt{\frac{n\sigma_z^2}{2R^2}}.$$

We now assume that $\hat{\mathbf{x}}$ is a local minimum of (3.26) with $\nu = \kappa_n(\sigma_z^2)$ such that $\rho > 0$. By the triangle inequality and Proposition 3.5.1,

$$d_H(\mathscr{A}(\hat{\mathbf{x}}), \mathscr{A}(\mathbf{x})) \leq d_H(\mathscr{A}(\hat{\mathbf{x}}), \bar{\mathscr{A}}(\mathbf{x})) + d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x})) \leq 2\sqrt{\frac{n\sigma_z^2}{2R^2}} + \sqrt{\frac{\log(1/\eta)}{2m}}.$$

Combining this with (3.29), we have that with probability at least $1 - 2\eta$,

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{R} \leq \frac{\sqrt{2}}{C_1}\sqrt{\frac{n\sigma_z^2}{R^2}} + \frac{1}{C_1}\sqrt{\frac{\log(1/\eta)}{2m}} + \frac{c_1}{C_1}\sqrt{\frac{n\log(18m) + \log(2/\eta)}{2m}}. \tag{3.31}$$

We next consider an alternative perspective on this model. Specifically, suppose that our observations are generated via

$$\bar{\mathscr{A}}_i(\mathbf{x}) = \mathscr{A}_i(\mathbf{x}_i') \quad \text{where} \quad \mathbf{x}_i' = \mathbf{x} + \mathbf{z}_i,$$

where $\mathbf{z}_i \sim \mathcal{N}(0, \sigma_z^2 I)$. Note that we can write this as

$$\mathscr{A}_i(\mathbf{x}_i') = \mathbf{a}_i^T(\mathbf{x} + \mathbf{z}_i) - \tau_i = \mathbf{a}_i^T\mathbf{x} - \tau_i + \mathbf{a}_i^T\mathbf{z}_i.$$

Since $\|\mathbf{a}_i\| = 1$, $\mathbf{a}_i^T\mathbf{z}_i \sim \mathcal{N}(0, \sigma_z^2)$, and thus this is equivalent to the model described above.

Thus, we can also interpret the above results as applying when each comparison is generated using a "misspecified" version of $\mathbf{x}$ which has been perturbed by Gaussian noise. Moreover, note that

$$\mathbb{E}\|\mathbf{x} - \mathbf{x}_i'\|^2 = \mathbb{E}\|\mathbf{z}_i\|^2 = n\sigma_z^2,$$

in which case we can also express the bound in (3.31) as

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{R} \leq \frac{\sqrt{2}}{C_1}\sqrt{\frac{\mathbb{E}\|\mathbf{x} - \mathbf{x}_i'\|^2}{R^2}} + \frac{1}{C_1}\sqrt{\frac{\log(1/\eta)}{2m}} + \frac{c_1}{C_1}\sqrt{\frac{n\log(18m) + \log(2/\eta)}{2m}}. \quad (3.32)$$

Thus, a small Gaussian perturbation of $\mathbf{x}$ in the comparisons will result in an increased recovery error roughly proportional to the (average) size of the perturbation.

Note that in establishing this result we apply Theorem 3.4.1, and so in contrast to our first noise model, here the result holds with high probability for a fixed $\mathbf{x}$ (as opposed to being uniform over all $\mathbf{x}$ for a single choice of $\mathscr{A}$).

Noise model 3..— In the third noise model, we assume the comparisons are generated according to

$$\bar{\mathscr{A}}(\mathbf{x}) = \mathscr{A}(\mathbf{x}'),$$

where $\mathbf{x}'$ represents an arbitrary perturbation of $\mathbf{x}$. Much like in the previous model, comparisons which are "decisive" are not likely to be affected by this kind of noise, while comparisons which are nearly even are quite likely to be affected. Unlike the previous model, our results here make no assumption on the distribution of the noise and will instead use the upper bound in Theorem 3.4.3 to establish a uniform guarantee that holds (with high probability) simultaneously for all choices of $\mathbf{x}$ (and $\mathbf{x}'$). Thus, in this model our guarantees are quite a bit stronger.

Specifically, we use the fact that from the upper bound of Theorem 3.4.3, with probability at least $1 - \eta$ we simultaneously have (3.27) and

$$d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x})) \leq C_2\frac{\|\mathbf{x} - \mathbf{x}'\|}{R} + c_2\zeta =: \kappa.$$

We again use (3.26) with $\nu = \kappa$ and Proposition 3.5.1 to produce an estimate $\hat{\mathbf{x}}$ satisfying $d_H(\mathscr{A}(\hat{\mathbf{x}}), \bar{\mathscr{A}}(\mathbf{x})) \leq \kappa$. Again using the triangle inequality, we have

$$d_H(\mathscr{A}(\hat{\mathbf{x}}), \mathscr{A}(\mathbf{x})) \leq d_H(\mathscr{A}(\hat{\mathbf{x}}), \bar{\mathscr{A}}(\mathbf{x})) + d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x})) \leq 2\kappa.$$

Combining this with (3.27) we have

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{R} \leq \frac{2\kappa + c_1\zeta}{C_1} = \frac{2C_2}{C_1}\frac{\|\mathbf{x} - \mathbf{x}'\|}{R} + \frac{c_1 + 2c_2}{C_1}\zeta.$$

Substituting in for ζ as in (3.29) yields

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{R} \leq \frac{2C_2}{C_1}\frac{\|\mathbf{x} - \mathbf{x}'\|}{R} + \frac{c_1 + 2c_2}{C_1}\sqrt{\frac{n\log(18m) + \log(2/\eta)}{2m}}. \tag{3.33}$$

Contrasting the result in (3.33) with that in (3.32), we note that up to constants, the results are essentially the same. This is perhaps somewhat surprising since (3.33) applies to arbitrary perturbations (as opposed to only Gaussian noise), and moreover, (3.33) is a uniform guarantee.

3.5.3 Adaptive estimation

Here we describe a simple extension to our previous (noiseless) theory and show that if we modify the mean and variance of the sampling distribution of items over a number of stages, we can localize *adaptively* and produce an estimate with many fewer comparisons than possible in a non-adaptive strategy. We assume t stages ($t = 1$ for the non-adaptive approach). At each stage $\ell \in [t]$ we will attempt to produce an estimate $\hat{\mathbf{x}}^\ell$ such that $\|\mathbf{x} - \hat{\mathbf{x}}^\ell\| \leq \epsilon_\ell$ where $\epsilon_\ell = R_\ell/2 = R2^{-\ell}$, then recentering to our previous estimate and dividing the problem radius in half. In stage ℓ, each $\mathbf{p}_i, \mathbf{q}_i \sim \mathcal{M}(\hat{\mathbf{x}}, 2R_\ell^2/n\mathbf{I})$. After t stages we will have $\|\mathbf{x} - \hat{\mathbf{x}}^t\| \leq R2^{-t} =: e_t$ with probability at least $1 - t\eta$.

Proposition 3.5.2. *Let $\epsilon_t, \eta > 0$ be given. Suppose that $\mathbf{x} \in \mathbb{B}_R^n$ and that m total comparisons are obtained following the adaptive scheme where*

$$m \geq 2C\log_2\left(\frac{2R}{\epsilon_t}\right)\left(n\log 2\sqrt{n} + \log\frac{1}{\eta}\right),$$

where C is a constant. Then with probability at least $1 - \log_2(2R/\epsilon_t)\eta$, for any estimate $\hat{\mathbf{x}}$ satisfying $\mathscr{A}(\hat{\mathbf{x}}) = \mathscr{A}(\mathbf{x})$,

$$\|\mathbf{x} - \hat{\mathbf{x}}\| \leq \epsilon_t.$$

Proof. The adaptive scheme uses $t = \lceil \log_2(R/\epsilon_t) \rceil \leq \log_2(2R/\epsilon_t)$ stages. Assume each stage is allocated m_ℓ comparisons. By Theorem 3.3.1, localization at each stage ℓ can be accomplished with high probability when

$$m_\ell \geq C \frac{R_\ell}{\epsilon_\ell} \left(n \log \frac{R_\ell \sqrt{n}}{\epsilon_\ell} + \log \frac{1}{\eta} \right) = 2C \left(n \log 2\sqrt{n} + \log \frac{1}{\eta} \right).$$

This condition is met by giving an equal number of comparisons to each stage, $m_\ell = \lfloor m/t \rfloor$. Each stage fails with probability η. By a union bound, the target localization fails with probability at most $t\eta$. Hence, localization succeeds with probability at least $1 - t\eta$. □

Proposition 3.5.2 implies $m_{\text{adapt}} \asymp (n \log n) \log_2(R/\epsilon_t)$ comparisons suffice to estimate $\mathbf{x}$ to within ϵ_t. This represents an exponential improvement in terms of number of total comparisons as a function of the target accuracy, ϵ_t, as compared to a lower bound on the number of required comparisons, $m_{\text{lower}} := 2nR/(e\epsilon_t)$ for *any* non-adaptive strategy (recall Theorem 3.3.3). Note that this result holds in the noise-free setting, but can easily be generalized to handle noisy settings via the approaches discussed above.

3.6 Simulations

In this section we perform a range of synthetic experiments to demonstrate our approach.

3.6.1 Effect of varying σ^2

In Fig. 3.2, we let $\mathbf{x} \in \mathbb{R}^2$ with $\|\mathbf{x}\| = R = 1$. We vary σ^2 and perform 1000 trials, each with $m = 50$ pairs of points drawn according to $\mathcal{N}(0, \sigma^2 I)$. To isolate the impact of σ^2, we consider the case where our observations are noise-free, and use (3.25) to recover $\hat{\mathbf{x}}$. As predicted by the theory, localization accuracy depends on the parameter σ, which controls the distribution of the hyperplane thresholds. Intuitively, if σ is too small, the hyperplane boundaries concentrate closer to the origin and do not localize points with large norm well.

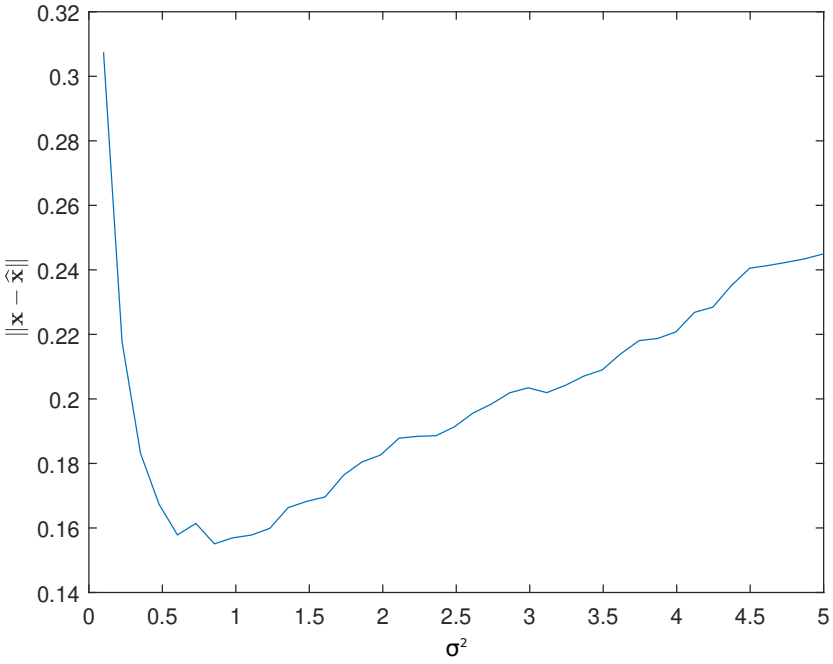


Figure 3.2: Mean error norm $\|\mathbf{x} - \hat{\mathbf{x}}\|$ as σ^2 varies.

On the other hand, if σ is too large, most hyperplanes lie far from the target $\mathbf{x}$. The sweet spot which allows uniform localization over the radius R ball exists around $\sigma^2 \approx 2R^2/n = 1$ here.

3.6.2 Effect of noise

Here we experiment with noise as discussed in Sections 3.4 and 3.5, and use the optimization program (3.26) for recovery. To approximately solve this non-convex problem, we use the linearization procedure described in Pérez-Cruz, Weston, Herrmann, and Schölkopf [149]. Specifically, over a number of iterations k, we repeatedly solve the sub-problem

$$\begin{aligned} \underset{\rho\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^m,\widehat{\mathbf{w}}^{(k)}\in\mathbb{R}^{n+1}}{\text{minimize}} \quad & -\nu\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & \bar{\mathscr{I}}_i(\mathbf{x})([\mathbf{a}_i^T, -\tau_i]\widehat{\mathbf{w}}^{(k)}) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \forall i \in [m], \\ & \widehat{\mathbf{w}}^{(k)T}\widetilde{\mathbf{w}}^{(k)} = 2 \end{aligned}$$

where we set $\widetilde{\mathbf{w}}^{(k+1)} \leftarrow \chi\widetilde{\mathbf{w}}^{(k)} + (1-\chi)\widehat{\mathbf{w}}^{(k)}$ with $\chi = 0.7$. After sufficient iterations, if $\widetilde{\mathbf{w}}^{(k)} \approx \widehat{\mathbf{w}}^{(k)}$ then (3.26) is approximately solved. This is a linear program and it can be easily

verified using the KKT conditions that $|\{i : \xi_i > 0\}| \leq m\nu$. Thus in practice, this property will always be satisfied after each iteration.

We also emphasize that the error bounds in Section 3.5 rely on the fact from Proposition 3.5.1 that $d_H(\mathscr{A}(\hat{\mathbf{x}}), \bar{\mathscr{A}}(\mathbf{x})) \leq \nu$, *provided that the solution results in a* $\rho > 0$. Unfortunately, we cannot guarantee that this will always be the case. Empirically, we have observed that given a certain noise level quantified by $d_H(\mathscr{A}(\mathbf{x}), \bar{\mathscr{A}}(\mathbf{x})) = \kappa$, we are more likely to observe $\rho \leq 0$ when we aggressively set $\nu = \kappa$. By increasing ν somewhat this becomes much less likely. As a rule of thumb, we set $\nu = 2\kappa$. We note that while in our context this choice is purely heuristic, it has some theoretical support in the ν-SVM literature (e.g., see Proposition 5 of [153]).

We consider the following noise models; *(i) Gaussian*, where we add pre-quantization Gaussian noise as in Section 3.4.1, *(ii) random*, where a uniform random $\nu/2$ fraction of comparisons are flipped, and *(iii) adversarial*, where we flip the $\nu/2$ fraction of comparisons whose hyperplane lie farthest from the ideal point. In each case, we set $n = 5$ and generate $m = 1000$ pairs of points and a random $\mathbf{x}$ with $\|\mathbf{x}\| = 0.7$. The mean and median recovery error $\|\hat{\mathbf{x}} - \mathbf{x}\|$ and the fraction of violated comparisons $d_H(\mathscr{A}(\hat{\mathbf{x}}), \mathscr{A}(\mathbf{x}))$ are plotted over 100 independent trials with varying number of comparison errors in Figs. 3.3–3.5. In both the Gaussian noise and uniform random comparison flipping cases, the actual fraction of comparison errors is on average much smaller than our target ν. This is also seen in the adversarial case (Fig. 3.5) for smaller levels of error. However, at a high fraction of error (greater than about 17%) the error (both in terms of Euclidean norm and fraction of incorrect comparisons) grows rapidly. This illustrates a limitation to the approach of using slack variables as a relaxation to the 0–1 loss. We mention that in this regime, the recovery approach of (3.26) frequently yields $\rho \leq 0$, to which our theory does not apply. This scenario, with a large number of erroneous comparisons, represents a very difficult situation in which any tractable recovery strategy would likely struggle. A possible direction for future work would be to make (3.26) more robust to such large outliers.

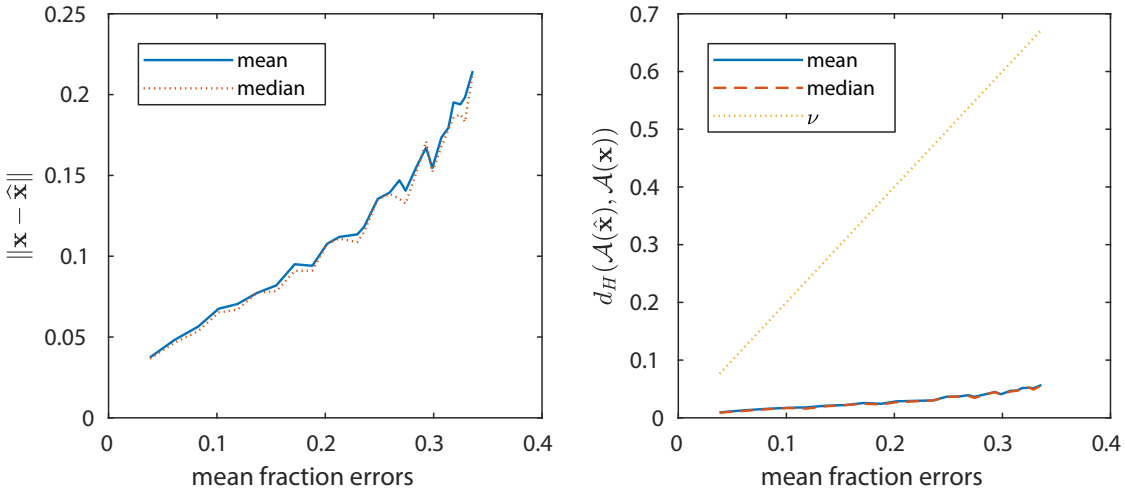


Figure 3.3: Estimation error and comparison errors when adding Gaussian noise.

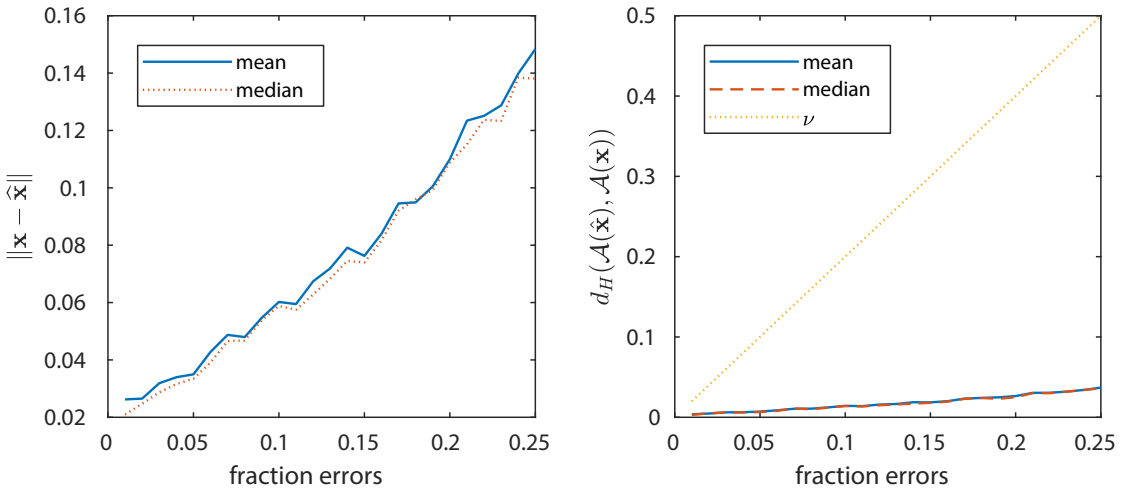


Figure 3.4: Estimation error and comparison errors with uniform random comparison errors.

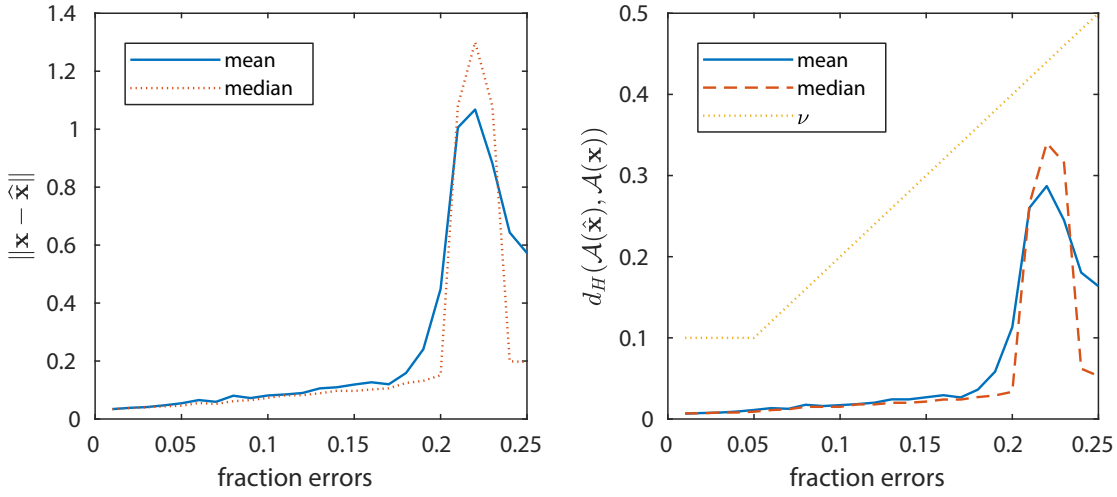


Figure 3.5: Estimation error and comparison errors when flipping the farthest comparisons.

3.6.3 Adaptive comparisons

In Fig. 3.6, we show the effect of varying levels of adaptivity, starting with the completely non-adaptive approach up to using 10 stages where we progressively re-center and re-scale the hyperplane offsets. In each case, we generate $\mathbf{x} \in \mathbb{R}^3$ where $\|\mathbf{x}\| = 0.75$ and choosing the direction randomly. The total number of comparisons are held fixed and are split as equally as possible among the number of stages (preferring earlier stages when rounding). We set $\sigma^2 = R = 1$ and plot the average over 700 independent trials. As the number of stages increases, performance worsens if the number of comparisons are kept small due to bad localization in the earlier stages. However, if the number of total comparisons is sufficiently large, an exponential improvement over non-adaptivity is possible.

3.6.4 Adaptive comparisons with a fixed non-Gaussian dataset

In Fig. 3.7, we demonstrate the effect of adaptively choosing item pairs from a fixed synthetic dataset over four stages versus choosing items non-adaptively, i.e., without attempting to estimate the signal during the comparison collection process. We first generated 10,000 items uniformly distributed inside the 3-dimensional unit ball and a vector $\mathbf{x} \in \mathbb{R}^3$ where $\|\mathbf{x}\| = 0.4$. In both cases, we generate pairs of Gaussian points and choose the items from the fixed dataset which lie closest to them. In the adaptive case over four stages, we progressively

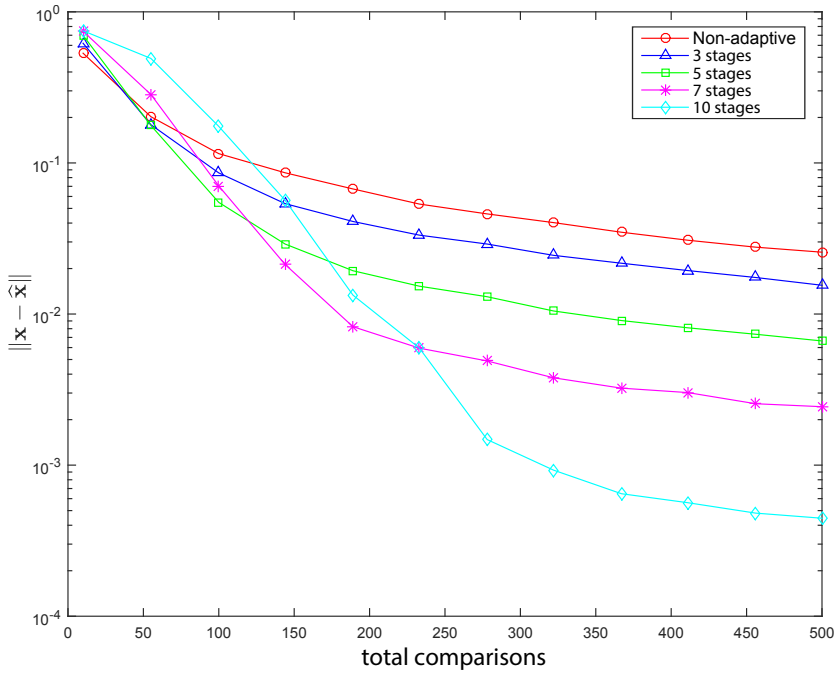


Figure 3.6: Mean error norm $\|\mathbf{x}-\hat{\mathbf{x}}\|$ versus total comparisons for a sequence of experiments with varying number of adaptive stages.

re-center and re-scale the generated points; the initial σ^2 is set to the variance of the dataset and is reduced dyadically after each stage. The total number of comparisons is held fixed and is split as equally as possible among the number of stages (preferring later stages when rounding). We plot the mean error over 200 independent-dataset trials.

3.7 Discussion

We have shown that given the ability to generate item pairs according to a Gaussian distribution with a particular variance, it is possible to estimate a point $\mathbf{x}$ satisfying $\|\mathbf{x}\| \leq R$ to within ϵ with roughly nR/ϵ paired comparisons (ignoring log factors). This procedure is also robust to a variety of forms of noise. If one is able to shift the distribution of the items drawn, adaptive estimation gives a substantial improvement over a non-adaptive strategy. To directly implement such a scheme, one would require the ability to generate items arbitrarily in $\mathbb{R}^n$. While there may be some cases where this is possible (e.g., in market testing of items where the features correspond to known quantities that can be manually manipulated, such as the amount of various ingredients in a food or beverage), in many of the settings

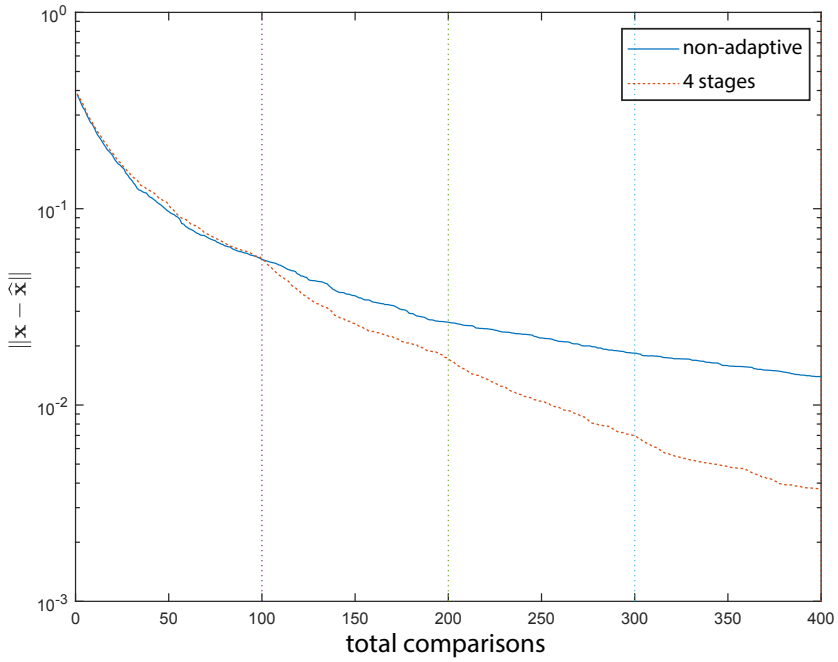


Figure 3.7: Mean error norm $\|\mathbf{x}-\hat{\mathbf{x}}\|$ versus total comparisons for nonadaptive and adaptive selection. Dotted lines denote stage boundaries.

considered by recommendation systems, the only items which can be compared belong to a fixed set of points. While our theory would still provide rough guidance as to how accurate of a localization is possible, many open questions in this setting remain. For instance, the algorithm itself needs to be adapted, as done in Section 3.6.4. Of course, there are many other ways that the adaptive scheme could be modified to account for this restriction. For example, one could use rejection sampling, so that although many candidate pairs would need to be drawn, only a fraction would actually need to be presented to and labeled by the user. We leave the exploration of such variations for future work.

3.8 Supporting lemmas

Lemma 3.8.1. *Let $b > a$ and let $L = \min\{|a|, |b|\}$ and $U = \max\{|a|, |b|\}$. Then if Φ and ϕ respectively denote the standard normal cumulative distribution function and probability distribution function, we have the bounds*

$$(b-a)\phi(U) \leq \Phi(b) - \Phi(a) \leq (b-a)\phi(L) \leq (b-a)\phi(0).$$

Proof. By the mean value theorem, we have for some $a < c < b$, $\Phi(b) - \Phi(a) = (b - a)\Phi'(c) = (b - a)\phi(c)$. Since $\phi(|x|)$ is monotonic decreasing, it is lower bounded by $\phi(U)$ and upper bounded by $\phi(L)$ (and also $\phi(0)$). □

Lemma 3.8.2. *Let* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. *Then,*

$$\int_{\mathbb{S}^{n-1}} |\mathbf{a}^T(\mathbf{x} - \mathbf{y})| \, \nu(\mathrm{d}\mathbf{a}) = \frac{2}{\sqrt{\pi}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})} \|\mathbf{x} - \mathbf{y}\|.$$

Proof. By spherical symmetry, we may assume $\Delta = \mathbf{x} - \mathbf{y} = [\epsilon, 0, \ldots, 0]$ for $\epsilon > 0$ without loss of generality. Then $\|\mathbf{x}-\mathbf{y}\| = \epsilon$ and $|\mathbf{a}^T(\mathbf{x}-\mathbf{y})| = a(1)\epsilon = \epsilon|\cos\theta|$, where $\cos^{-1}(a(1)) = \theta \in [0, \pi]$. We will use the fact [154]:

$$\int_0^{\frac{\pi}{2}} \cos^{\mu-1}\theta \sin^{\omega-1}\theta \, \mathrm{d}\theta = \frac{1}{2} B\left(\frac{\mu}{2}, \frac{\omega}{2}\right) = \frac{1}{2} \frac{\Gamma(\mu/2)\Gamma(\omega/2)}{\Gamma((\mu+\omega)/2)}.$$

Integrating $|\cos\theta|$ in the first spherical coordinate, since the integrand is symmetric about $\frac{\pi}{2}$,

$$\int_0^{\pi} |\cos\theta| \sin^{n-2}\theta \, \mathrm{d}\theta = 2\int_0^{\pi/2} \cos\theta \sin^{n-2}\theta \, \mathrm{d}\theta = \frac{\Gamma(1)\Gamma(\frac{n-1}{2})}{\Gamma(1 + \frac{n-1}{2})} = \frac{2}{n-1}.$$

Then with the appropriate normalization, we have (using $\Gamma(1/2) = \sqrt{\pi}$)

$$\begin{aligned}\int_{S^{n-1}} |a^T(\mathbf{x} - \mathbf{y})| \, \nu(\mathrm{d}a) &= \left(\int_0^{\pi} \sin^{n-2}\theta \, \mathrm{d}\theta\right)^{-1} \int_0^{\pi} \epsilon|\cos\theta| \sin^{n-2}\theta \, \mathrm{d}\theta \\ &= \epsilon \left(\frac{\Gamma(\frac{1}{2})\Gamma(\frac{n-1}{2})}{\Gamma(\frac{1}{2} + \frac{n-1}{2})}\right)^{-1} \frac{2}{n-1} = \frac{2}{\sqrt{\pi}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})} \|\mathbf{x} - \mathbf{y}\|.\end{aligned}$$

□

3.9 Integral calculations for Lemma 3.4.2

First, we give an expression for κ_n for all cases $n \geq 2$, expanding upon that given in Theorem 3.4.1 and Lemma 3.4.2. We have

$$\kappa_n(\sigma_z^2) := \begin{cases} \frac{1}{2}\sqrt{\frac{\sigma_z^2}{\sigma_z^2+R^2}} & n=2 \\ \min\left\{ \sqrt{\frac{\sigma_z^2}{\sigma_z^2+2R^2/3}},\ \sqrt{\frac{\pi}{2}}\frac{\sigma_z}{\|\mathbf{x}\|} \right\} & n=3 \\ \sqrt{\frac{\sigma_z^2}{\sigma_z^2+2R^2/n+4\|\mathbf{x}\|^2/n}} & n \geq 4. \end{cases}$$

Below we derive this expression for the cases $n = 2$, $n = 3$, and $n \geq 4$.

3.9.1 Case $n = 2$

For the special case $n = 2$, $d_i = \cos\theta_i$ where $\theta_i \in [-\pi, \pi]$ is distributed uniformly. In this case, (3.21) can be re-written as

$$\begin{aligned} \mathbb{P}[q_i\bar{q}_i < 0] &\leq \frac{1}{2R}\sqrt{\frac{2}{\pi}} \int_{-\pi/2}^{\pi/2}\int_{-\infty}^{\infty} \frac{1}{2\pi}\exp\left(-\frac{(\|\mathbf{x}\|\cos\theta_i - \tau_i)^2}{2\sigma_z^2} - \frac{\tau_i^2}{2R^2}\right) \mathrm{d}\tau_i\,\mathrm{d}\theta_i \\ &= \frac{1}{\pi R}\sqrt{\frac{1}{2\pi}} \int_{0}^{\pi/2}\int_{-\infty}^{\infty} \exp\left(-\frac{(\|\mathbf{x}\|\cos\theta_i - \tau_i)^2}{2\sigma_z^2} - \frac{\tau_i^2}{2R^2}\right) \mathrm{d}\tau_i\,\mathrm{d}\theta_i. \end{aligned}$$

Expanding and setting α, β, and γ appropriately,

$$\begin{aligned} \mathbb{P}[q_i\bar{q}_i < 0] &\leq \frac{1}{\pi R}\sqrt{\frac{1}{2\pi}} \int_{0}^{\pi/2}\int_{-\infty}^{\infty} \exp\left(-\frac{\|\mathbf{x}\|^2\cos^2\theta_i}{2\sigma_z^2} + \frac{2\|\mathbf{x}\|\tau_i\cos\theta_i}{2\sigma_z^2} - \frac{\tau_i^2}{2\sigma_z^2} - \frac{\tau_i^2}{2R^2}\right) \mathrm{d}\tau_i\,\mathrm{d}\theta_i \\ &= \frac{1}{\pi R}\sqrt{\frac{1}{2\pi}} \int_{0}^{\pi/2}\int_{-\infty}^{\infty} \exp\left(-\gamma\cos^2\theta_i + \beta\tau_i\cos\theta_i - \alpha\tau_i^2\right) \mathrm{d}\tau_i\,\mathrm{d}\theta_i. \end{aligned}$$

Completing the square for τ_i,

$$\begin{aligned} \mathbb{P}[q_i\bar{q}_i < 0] &= \frac{1}{\pi R}\sqrt{\frac{1}{2\pi}} \int_{0}^{\pi/2}\int_{-\infty}^{\infty} \exp\left(-\alpha\left(\tau_i + \frac{\beta\cos\theta_i}{2\alpha}\right)^2 + \frac{(\beta\cos\theta_i)^2}{4\alpha} - \gamma\cos^2\theta_i\right) \mathrm{d}\tau_i\,\mathrm{d}\theta_i \\ &= \frac{1}{\pi R}\sqrt{\frac{1}{2\pi}} \int_{0}^{\pi/2} \sqrt{\frac{\pi}{\alpha}}\exp\left(-\left(\gamma - \frac{\beta^2}{4\alpha}\right)\cos^2\theta_i\right) \mathrm{d}\theta_i \\ &= \frac{\pi}{2\pi R}\sqrt{\frac{1}{2\alpha}}\exp\left(-\frac{1}{2}\left(\gamma - \frac{\beta^2}{4\alpha}\right)\right) I_0\left(\frac{1}{2}\left(\gamma - \frac{\beta^2}{4\alpha}\right)\right), \end{aligned}$$

where $I_0(\cdot)$ denotes the modified Bessel function of the first kind. Since $\exp(-t)I_0(t) < 1$, by plugging back in for α we obtain

$$\mathbb{P}[q_i\bar{q}_i < 0] \leq \frac{1}{2R}\sqrt{\frac{1}{2\alpha}} = \frac{1}{2\sqrt{2}R}\sqrt{\frac{1}{\frac{1}{2\sigma_z^2}+\frac{1}{2R^2}}} = \frac{1}{2}\sqrt{\frac{\sigma_z^2}{\sigma_z^2+R^2}}.$$

We also note that since since $\exp(-t)I_0(t) < 1/\sqrt{\pi t}$, we can obtain the bound $\mathbb{P}[q_i\bar{q}_i < 0] \leq \frac{1}{\sqrt{\pi}}\frac{\sigma_z}{\|\mathbf{x}\|}$, but one can show that the previous bound will dominate this whenever $\|\mathbf{x}\| \leq R$.

3.9.2 Case $n = 3$

For the case $n = 3$, $d_i \sim [-1, 1]$ is itself distributed uniformly. In this case we have

$$\mathbb{P}[q_i\bar{q}_i < 0] \leq \frac{1}{2R}\sqrt{\frac{3}{\pi}}\int_0^1\int_{-\infty}^{\infty}\exp\left(-\frac{(d_i\|\mathbf{x}\|-\tau_i)^2}{2\sigma_z^2}-\frac{3\tau_i^2}{4R^2}\right)\mathrm{d}\tau_i\,\mathrm{d}d_i.$$

Expanding and setting α, β, and γ appropriately,

$$\begin{aligned}\mathbb{P}[q_i\bar{q}_i < 0] &\leq \frac{1}{2R}\sqrt{\frac{3}{\pi}}\int_0^1\int_{-\infty}^{\infty}\exp\left(-\frac{d_i^2\|\mathbf{x}\|^2}{2\sigma_z^2}+\frac{2d_i\|\mathbf{x}\|\tau_i}{2\sigma_z^2}-\frac{\tau_i^2}{2\sigma_z^2}-\frac{3\tau_i^2}{4R^2}\right)\mathrm{d}\tau_i\,\mathrm{d}d_i \\ &= \frac{1}{2R}\sqrt{\frac{3}{\pi}}\int_0^1\int_{-\infty}^{\infty}\exp\left(-\gamma d_i^2+\beta d_i\tau_i-\alpha\tau_i^2\right)\mathrm{d}\tau_i\,\mathrm{d}d_i.\end{aligned}$$

Completing the square for τ_i,

$$\begin{aligned}\mathbb{P}[q_i\bar{q}_i < 0] &= \frac{1}{2R}\sqrt{\frac{3}{\pi}}\int_0^1\int_{-\infty}^{\infty}\exp\left(-\alpha\left(\tau_i+\frac{d_i\beta}{2\alpha}\right)^2+\frac{(d_i\beta)^2}{4\alpha}-\gamma d_i\right)\mathrm{d}\tau_i\,\mathrm{d}d_i \\ &= \frac{1}{2R}\sqrt{\frac{3}{\pi}}\int_0^1\sqrt{\frac{\pi}{\alpha}}\exp\left(-d_i^2\left(\gamma-\frac{\beta^2}{4\alpha}\right)\right)\mathrm{d}d_i \\ &= \frac{1}{2R}\sqrt{\frac{3}{\alpha}}\frac{\sqrt{\pi}}{2}\frac{\operatorname{erf}\left(\sqrt{\gamma-\beta^2/4\alpha}\right)}{\sqrt{\gamma-\beta^2/4\alpha}}.\end{aligned}$$

Since $\operatorname{erf}(t)/t \leq 2/\sqrt{\pi}$, by plugging back in for α we obtain

$$\mathbb{P}[q_i\bar{q}_i < 0] \leq \frac{1}{2R}\sqrt{\frac{3}{\frac{1}{2\sigma_z^2}+\frac{3}{4R^2}}} = \sqrt{\frac{\sigma_z^2}{\sigma_z^2+2R^2/3}}.$$

Additionally, since $\operatorname{erf}(t) \leq 1$,

$$\begin{aligned}
\mathbb{P}[q_i \bar{q}_i < 0] &\leq \frac{\sqrt{3\pi}}{4R} \left(\gamma\alpha - \beta^2/4\right)^{-1/2} \\
&= \frac{\sqrt{3\pi}}{4R} \left(\frac{\|\mathbf{x}\|^2}{2\sigma_z^2} \left(\frac{1}{2\sigma_z^2} + \frac{3}{4R^2} \right) - \frac{\|\mathbf{x}\|^2}{4\sigma_z^4} \right)^{-1/2} \\
&= \frac{\sqrt{3\pi}}{4R} \left(\frac{3\|\mathbf{x}\|^2}{8\sigma_z^2 R^2} \right)^{-1/2} \\
&= \sqrt{\frac{\pi}{2}} \frac{\sigma_z}{\|\mathbf{x}\|},
\end{aligned}$$

which can be tighter when σ_z is small and $\|\mathbf{x}\|$ is large.

3.9.3 Case $n \geq 4$

Combining (3.21) with our upper bound (3.22) on $f_d(d_i)$, we obtain

$$\mathbb{P}[q_i \bar{q}_i < 0] \leq \frac{n}{2\sqrt{2}\pi R} \int_0^1 \int_{-\infty}^{\infty} \exp\left(-\frac{(d_i\|\mathbf{x}\| - \tau_i)^2}{2\sigma_z^2} - \frac{n\tau_i^2}{4R^2} - \frac{nd_i^2}{8} \right) \mathrm{d}\tau_i \, \mathrm{d}d_i.$$

Expanding and setting α, β, and γ appropriately,

$$\begin{aligned}
\mathbb{P}[q_i \bar{q}_i < 0] &\leq \frac{n}{2\sqrt{2}\pi R} \int_0^1 \int_{-\infty}^{\infty} \exp\left(-d_i^2 \left(\frac{\|\mathbf{x}\|^2}{2\sigma_z^2} + \frac{n}{8} \right) + \frac{2d_i\|\mathbf{x}\|\tau_i}{2\sigma_z^2} - \frac{\tau_i^2}{2\sigma_z^2} - \frac{n\tau_i^2}{4R^2} \right) \mathrm{d}\tau_i \, \mathrm{d}d_i \\
&= \frac{n}{2\sqrt{2}\pi R} \int_0^1 \int_{-\infty}^{\infty} \exp\Big(-\gamma d_i^2 + \beta d_i \tau_i - \alpha \tau_i^2 \Big) \, \mathrm{d}\tau_i \, \mathrm{d}d_i.
\end{aligned}$$

Completing the square for τ_i,

$$\begin{aligned}
\mathbb{P}[q_i \bar{q}_i < 0] &= \frac{n}{2\sqrt{2}\pi R} \int_0^1 \int_{-\infty}^{\infty} \exp\left(-\alpha \left(\tau_i - \frac{d_i\beta}{2\alpha} \right)^2 + \frac{(d_i\beta)^2}{4\alpha} - \gamma d_i^2 \right) \mathrm{d}\tau_i \, \mathrm{d}d_i \\
&= \frac{n}{2\sqrt{2}\pi R} \int_0^1 \sqrt{\frac{\pi}{\alpha}} \exp\left(-d_i^2 \left(\gamma - \frac{\beta^2}{4\alpha} \right) \right) \mathrm{d}d_i \\
&= \frac{n}{2\sqrt{2\pi\alpha}R} \frac{\sqrt{\pi}}{2} \frac{\operatorname{erf}\left(\sqrt{\gamma - \beta^2/4\alpha} \right)}{\sqrt{\gamma - \beta^2/4\alpha}}.
\end{aligned}$$

Since $\operatorname{erf}(t) \leq 1$, we have

$$\begin{aligned}
\mathbb{P}[q_i \bar{q}_i < 0] =& \leq \frac{n}{4\sqrt{2}R} \left(\gamma\alpha - \beta^2/4\right)^{-1/2} \\
&= \frac{n}{4\sqrt{2}R} \left(\left(\frac{\|\mathbf{x}\|^2}{2\sigma_z^2} + \frac{n}{8} \right) \left(\frac{1}{2\sigma_z^2} + \frac{n}{4R^2} \right) - \frac{\|\mathbf{x}\|^2}{4\sigma_z^4} \right)^{-1/2} \\
&= \left(\frac{32R^2}{n^2} \left(\frac{\|\mathbf{x}\|^2 n}{8\sigma_z^2 R^2} + \frac{n}{16\sigma_z^2} + \frac{n^2}{32R^2} \right) \right)^{-1/2} \\
&= \sqrt{\frac{\sigma_z^2}{\sigma_z^2 + 2R^2/n + 4\|\mathbf{x}\|^2/n}}.
\end{aligned}$$

We also note that since $\operatorname{erf}(t)/t \leq 2/\sqrt{\pi}$, it is also possible to obtain the bound

$$\mathbb{P}[q_i \bar{q}_i < 0] \leq \sqrt{\frac{n}{2\pi}} \sqrt{\frac{\sigma_z^2}{\sigma_z^2 + 2R^2/n}}.$$

However, this bound can only be tighter when $\|\mathbf{x}\|$ is small and when $\frac{n}{2\pi} < 1$ (i.e., for $n \leq 6$). Given this narrow range of applicability, we omit this from the formal statement of the result.

CHAPTER 4

ACTIVE EMBEDDING SEARCH VIA NOISY PAIRED COMPARISONS

In this chapter we consider the paired comparison setting introduced in Chapter 3. Note that the notation in this chapter differs slightly from that of Chapter 3. Specifically, here $\mathbf{w} \in \mathbb{R}^d$ represents the signal to be estimated and d is the dimension of space. Capital $\mathbf{W}$ is used when referring to the random variable which represents our current knowledge of $\mathbf{w}$. As before, we wish to estimate a vector $\mathbf{w}$ from paired comparisons of the form "is $\mathbf{w}$ more similar to item $\mathbf{p}$ or to item $\mathbf{q}$?" In such tasks, queries can be extremely costly, thus we aim to choose pairs *actively* given the results of previous comparisons. By imposing a simple probabilistic model on responses, we show that adaptive selection can lead to better estimation with dramatically fewer queries. We greedily choose pairs which maximize the information gain and then develop two heuristics which maximize a lower bound on mutual information and are simpler to analyze and compute, respectively. We give bounds on the expected number of queries required to achieve a certain performance, and we validate our approach using simulated responses from a real-world dataset.

4.1 Introduction

We consider the problem of user preference learning, where we have a set of *items* (e.g., movies, music, or food) embedded in a Euclidean space and would like to determine the preferences of *users* with respect to these items such that users with similar preferences are assigned similar embedding coordinates. We will do this using the method of paired comparisons, where during each interaction a user chooses which one of two given items they prefer [128]. For instance, to find a particular food that a person would like to eat, we might ask them a number of queries in which they select one of two food items as more preferable in taste. From this information, we find an estimate of the preference point w. This

Note.—this chapter is joint work with Greg Canal, Mark Davenport, and Chris Rozell and has been submitted for publication.

recovered preference point can be used in various tasks, for instance in the recommendation of nearby items or clustering of users with similar preferences. We refer to this process as *pairwise search*, and a key goal of ours is to choose these items *actively* and demonstrate the advantage over non-adaptive selection of items.

More specifically, given N items to select from, there are $O(N^2)$ possible queries. Exhaustively querying all such pairs is not only prohibitively expensive for large datasets, but is also unnecessary. Intuitively, the set of all possible paired comparisons contains queries rendered obvious by the accumulation of evidence about the user's point, as well as ambiguous queries that provide no information due to their unreliability in the context of a model that accounts for noise in the comparison process. The main aim of this work is to design a query selection algorithm that only selects *informative* pairs in identifying a user's preference point. We achieve this by selecting pairs that aim to maximize the mutual information between the user's response and their true, unknown preference point.

Our approach relies on the use of mutual information to provide theoretical insights concerning the user point estimation problem and to provide lower bounds on the estimation error achievable by any query strategy. We then present upper and lower bounds on performance of a heuristic strategy that maximizes a lower bound on mutual information. Since estimating the mutual information of each pair in a pool can be computationally expensive, we motivate the use of a second, computationally cheaper heuristic that also maximizes a lower bound on mutual information while still performing comparably to naïve information maximization. We evaluate our methods against randomly selected measurements on a test dataset and demonstrate the benefits of our informative query strategy. To the best of our knowledge, our work is the first attempt to search a low-dimensional embedding for a *continuous* point via paired comparisons with *noisy* response models.

4.2 Background

4.2.1 Observation model

Our goal in this chapter is to estimate a user's preference point w with respect to a low-dimensional embedding of items. Specifically, we suppose user preferences can be captured via an *ideal point model* in which each item and user is represented using a common set of parameters. For simplicity, we assume that users and items are represented by points in $\mathbb{R}^d$ and that a user's overall preference for a particular item decreases with the distance between that item and the user's ideal point w. This means any item placed exactly at the user would be considered "ideal" and would be the most preferred over all other items. Although this model can be applied to the situation where a particular item is sought, in general we do *not* assume the user's location w to be co-located with any item.

The low-dimensional embedding of the items can be constructed through a training set of triplet comparisons (paired comparisons regarding similarity to a third reference item) using one of several standard non-metric embedding methods such as the Crowd Kernel Learning [155] or Stochastic Triplet Embedding methods [156]. In this study, we assume that such an embedding is given, presumably acquired through a large set of crowd-sourced training triplet comparisons. We do not consider this training set to be part of the learning cost in measuring a search algorithm's efficiency, since our focus here is on efficiently choosing paired comparisons to search an *existing* embedding.

In this work, we assume a noisy observation model where the probability of a user located at w choosing item p over item q is modeled using

$$\mathbb{P}(\mathbf{p} \prec \mathbf{q}) = f(k_{pq}(\|\mathbf{w} - \mathbf{q}\|^2 - \|\mathbf{w} - \mathbf{p}\|^2)), \tag{4.1}$$

where $\mathbf{p} \prec \mathbf{q}$ denotes "item $\mathbf{p}$ is preferred to item $\mathbf{q}$," $f(x) = 1/(1 + e^{-x})$ is the logistic function, and k_{pq} is the pair's *noise constant*. This type of logistic noise model is common in psychometric literature and our model bears similarity to the Bradley–Terry model [157].

Note that (4.1) can also be written as

$$\begin{aligned}\mathbb{P}(\mathbf{p} \prec \mathbf{q}) &= f(k_{pq}(\|\mathbf{w}-\mathbf{q}\|^2 - \|\mathbf{w}-\mathbf{p}\|^2)) \\ &= f(k_{pq}(2(\mathbf{p}-\mathbf{q})^T\mathbf{w} - \|\mathbf{p}\|^2 + \|\mathbf{q}\|^2)) \\ &= f(k_{pq}(\mathbf{a}^T\mathbf{w} - b)),\end{aligned}$$

where $\mathbf{a} = 2(\mathbf{p}-\mathbf{q})$ and $\mathbf{b} = \|\mathbf{p}\|^2 + \|\mathbf{q}\|^2$ encode the weights and threshold of a hyperplane bisecting items $\mathbf{p}$ and $\mathbf{q}$. k_{pq} represents roughly the signal-to-noise ratio of a particular measurement, which is dependent on the values of $\mathbf{p}$ and $\mathbf{q}$. After observing the results of a number of such queries, the response model in (4.1) for each query can be multiplied to form a posterior belief about the location of $\mathbf{w}$, as depicted in Figure 4.1.

Note that we allow the noise constant k_{pq} to differ for each item pair to allow for differing user behavior depending on the geometry of the items being compared. When $k_{pq} \to \infty$, this supposes a user's selection is made with complete certainty and cannot be erroneous. Conversely, $k_{pq} = 0$ corresponds to choosing items randomly with probability 1/2. Varying k_{pq} allows for differing reliability when items are far apart versus when they are close. Some concrete examples for setting this parameter are:

$$\text{constant} : k_{pq}^{(1)} = k_0, \tag{K1}$$

$$\text{normalized} : k_{pq}^{(2)} = k_0\|\mathbf{a}\|^{-1} = k_0\|2(\mathbf{p}-\mathbf{q})\|^{-1}, \tag{K2}$$

$$\begin{aligned}\text{decaying} : k_{pq}^{(3)} &= k_0 \exp(-\|\mathbf{a}\|) \\ &= k_0 \exp(-\|2(\mathbf{p}-\mathbf{q})\|).\end{aligned} \tag{K3}$$

4.2.2 Related work

There is a large body of work investigating preference learning and ranking via paired comparisons. Many of these works aim to construct embeddings, train classifiers, or learn rankings over items (for instance, [125, 137, 158, 159]). In contrast to these works, we do not intend to build an embedding of items, but instead quickly locate an individual user.

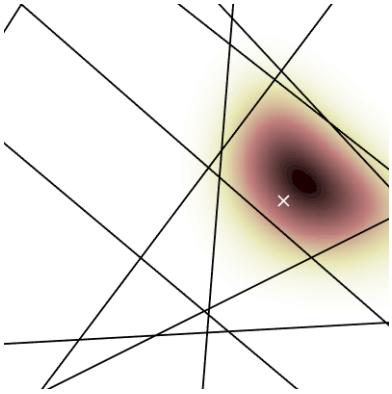


Figure 4.1: Paired comparisons between items can be thought of as a set of noisy hyperplane queries. In the high-fidelity case, the set of hyperplanes uniquely identifies a convex region of $\mathbb{R}^d$. More generally, we have a posterior distribution which only approximates the shape of the ideal cell around the true user point, depicted here with a cross.

For instance, noisy paired comparisons are utilized to construct triplet embeddings by repeatedly selecting which of two items is more similar to a reference item. While this technique has been used to construct low-dimensional item embeddings [155] (which can subsequently be searched with our technique), it does not directly explore the user point search problem. In a classification setting such as in [160], noisy paired comparisons can be used to estimate the parameters of a hyperplane decision surface, but this statistical estimation problem is inherently different from the one studied here since rather than having a single response model, we have a separate response model for each pair in a sequence of queries. Our model also bears similarity to standard setup in logistic regression (see e.g., [161]) but is subtly different due to the presence of "offsets" b, whereas logistic regression includes only the linear weights a. These offsets are necessary to model paired comparisons between items and adds considerable complexity to the geometry of the problem.

More generally, the presence of distinct hyperplane offsets precludes the low dimensional search setting studied here from being framed as the commonly used Bradley–Terry model where the probability model for a query is equal to the ratio of score functions over items. This is a common model in the active ranking setting, where adaptive comparisons aim to learn a score function or top-K ranking over a set of items [162, 163]. The difference

in our work is that our model cannot be decomposed into a ratio of score functions, nor are we exclusively searching for an item within a given dataset [124].

Similarly, in the setting of multi-armed or dueling bandits the goal is to find the "top reward" item in a dataset via adaptive querying [164]. As in the active ranking setting, this problem specifically seeks an item within a given dataset (without exploiting any underlying geometry) rather than estimating a continuous user point. Furthermore, in the bandit setting the queries may or may not involve paired comparisons.

The work that is most similar to the problem studied here is [135], which aims to determine the ranking of an item based on paired comparisons with other items in a low-dimensional embedding. A crucial difference between [135] and our work is that we are not merely interested in learning a rank-ordering among items, but instead seek to estimate a continuous user point. In contrast to the ranking of items alone, knowledge of the underlying user parameter in our model gives a us a stronger understanding of the user's behavior which can be used, for instance, to predict preferences for new items, as a component in an approach which alternates the task of user and item coordinate estimation, or as a step towards user clustering or other goals. Furthermore, the utilization of a continuous geometric model allows us to potentially generate queries by *synthesizing* points in the low-dimensional space.

4.3 Query selection

4.3.1 Information theoretic framework

Let $\mathbf{W} \in \mathbb{R}^d$ denote a random vector encoding the user preference point, assumed for the sake of analysis to be drawn from a uniform distribution over the hypercube $[-\frac{1}{2}, \frac{1}{2}]^d$. Unless noted otherwise, we denote random variables with uppercase letters, and specific realizations with lowercase letters. Let $Y_i \in \{0, 1\}$ denote the binary response to the i^{th} paired comparison involving items $\mathbf{p}_i$ and $\mathbf{q}_i$, with $Y_i = 0$ (resp. 1) indicating a user preference for $\mathbf{p}_i$ (resp. $\mathbf{q}_i$). After i queries, we have the vector of responses $\mathbf{Y}^i = \{Y_1, Y_2, \dots Y_i\}$.

Denoting the prior density as $p_0(\mathbf{w})$, after i queries we have a posterior density of

$$p_i(\mathbf{w}) \equiv p(\mathbf{w}|\mathbf{Y}^i) = \frac{p(Y_i|\mathbf{w})p(\mathbf{w}|\mathbf{Y}^{i-1})}{p(Y_i|\mathbf{Y}^{i-1})}. \tag{4.2}$$

The logistic response model used here for $p(Y_j|\mathbf{w})$ belongs to the class of *log-concave* distributions. Such distributions have probability density functions $f(\mathbf{w})$ satisfying $f(\alpha w_1 + (1-\alpha)\mathbf{w}_2) \geq f(\mathbf{w}_1)^\alpha f(\mathbf{w}_2)^{1-\alpha}$ for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ and $0 \leq \alpha \leq 1$. The uniform prior on $\mathbf{W}$ over the unit hypercube is also log-concave. Consequently, since products of log-concave functions are also log-concave, we have that the posterior density given in (4.2) is log-concave [165].

After i queries, the posterior $p_i(\mathbf{w})$ is used to generate a user point estimate $\widehat{\mathbf{W}}_i$. We denote the mean-squared error for this estimate by $\text{MSE}_i = \mathbb{E}_{\mathbf{W}|\mathbf{Y}^i}[\|\mathbf{W} - \widehat{\mathbf{W}}_i\|_2^2]$. At each step, we desire to utilize knowledge from previous queries to select a pair $(\mathbf{p}_i, \mathbf{q}_i)$ such that MSE_{i+1} is minimized after the user responds. However, such a scheme would require both updating the posterior distribution and estimating MSE_{i+1} for each possible response over all pairwise queries in a pool, which is very computationally expensive since in the model studied here each such evaluation requires a new set of Monte Carlo samples from the posterior (since there is no closed-form solution for MSE_{i+1} under our model). This is suboptimal for adaptive querying settings where typically data sets are large (resulting in a large number of candidate pairwise queries) and queries need to be selected in or close to real-time.

Instead, consider the covariance matrix of the user point posterior after i queries, denoted as $\mathbf{\Sigma}_{\mathbf{W}|\mathbf{Y}^i} = \mathbb{E}[(\mathbf{W} - \mathbb{E}[\mathbf{W}|\mathbf{Y}^i])(\mathbf{W} - \mathbb{E}[\mathbf{W}|\mathbf{Y}^i])^T|\mathbf{Y}^i]$. For a mean estimator of the user point given by $\widehat{\mathbf{W}}_i = \mathbb{E}[\mathbf{W}|\mathbf{Y}^i]$, which is the minimum mean-squared error (MMSE) estimator, we have

$$\text{MSE}_i = \text{Tr}(\mathbf{\Sigma}_{\mathbf{W}|\mathbf{Y}^i}) \geq d|\mathbf{\Sigma}_{\mathbf{W}|\mathbf{Y}^i}|^{\frac{1}{d}} \tag{4.3}$$

where the last inequality is from the arithmetic-geometric mean inequality (AM–GM) [105]. This implies that a necessary condition for minimizing MSE is to minimize the determinant of the posterior covariance matrix (denoted by $|\mathbf{\Sigma}_{\mathbf{W}|\mathbf{Y}^i}|$), which we refer to as the *posterior*

volume. Unfortunately, for the same reasons described above, selecting queries that directly minimize posterior volume is too computationally expensive to be useful in practice.

However, by utilizing statistical tools from information theory, we can select queries that approximately minimize posterior volume (and hence MSE) while still selecting queries in a computationally feasible manner. Furthermore, an information theoretic approach provides convenient analytical tools which we use to provide performance guarantees for the query selection methods we present.

Towards this end, we define the *posterior entropy* as the differential entropy of the posterior after i queries:

$$h_i(\mathbf{W}) \equiv h(\mathbf{W}|\mathbf{y}^i) = \int_w p_i(\mathbf{w}) \log_2(1/p_i(\mathbf{w}))dw. \tag{4.4}$$

As we show in the following lemma, posterior entropy is both upper and lower bounded by a monotonically increasing function of posterior volume, implying that minimizing posterior entropy is both a necessary and sufficient condition for minimizing the posterior volume, and hence a necessary condition for minimizing MSE. The proof of this lemma and all subsequent results are provided in the supplementary material.

Lemma 4.3.1. *For a log-concave posterior distribution $p(\mathbf{W}|\mathbf{y}^i)$ in d dimensions,*

$$\frac{d}{2}\log_2 \frac{2|\mathbf{\Sigma}_{\mathbf{W}|\mathbf{Y}^i}|^{\frac{1}{d}}}{c_d} - d \leq h_i(\mathbf{W}) \leq \frac{d}{2}\log_2(2\pi e|\mathbf{\Sigma}_{\mathbf{W}|\mathbf{Y}^i}|^{\frac{1}{d}}),$$

where $c_d = (e^2 d^2)(4\sqrt{2}(d+2))$.

This relationship between MSE, posterior volume, and posterior entropy minimization suggests a strategy of selecting queries that minimize the differential entropy of the posterior distribution after the query. Since the actual user response is unknown at the time of query selection, we seek to minimize the *expected* posterior entropy after a response is made, i.e., $\mathbb{E}_{Y_{i+1}}[h_{i+1}(\mathbf{W})|\mathbf{y}^i]$. Using a standard result from information theory, we have $\mathbb{E}_{Y_{i+1}}[h_{i+1}(\mathbf{W})|\mathbf{y}^i] = h_i(\mathbf{W}) - I(\mathbf{W}; Y_{i+1}|\mathbf{y}^i)$, where $I(\mathbf{W}; Y_{i+1}|\mathbf{y}^i)$ is the *mutual information*

between a query and the user response given previous responses [166].

Examining this identity, we observe that selecting queries that minimize the expected posterior entropy is equivalent to selecting queries that *maximize the mutual information* between the user point and the user response response.

In this setting, it is generally difficult to obtain sharp performance bounds for query selection via mutual information maximization. Instead, we use information theoretic tools along with Lemma 4.3.1 to provide a lower bound on MSE for *any* estimator and query selection scheme in a manner similar to estimation lower bounds presented in [166]:

Theorem 4.3.2. *For any user point estimator given by $\widehat{\mathbf{W}}_i$ after i queries, the MSE (averaged over user points and query responses) for any query selection strategy is lower bounded by*

$$\mathbb{E}_{\mathbf{W},\mathbf{Y}^i}\|\mathbf{W}-\widehat{\mathbf{W}}_i\|_2^2 \geq \frac{d2^{-2\frac{i}{d}}}{2\pi e}.$$

Using the symmetry of mutual information [166], we can write

$$I(\mathbf{W};Y_{i+1}|\mathbf{y}^i) = H(Y_{i+1}|\mathbf{y}^i) - H(Y_{i+1}|\mathbf{W},\mathbf{y}^i) \tag{4.5}$$

where

$$H(Y_{i+1}|\mathbf{y}^i) = \sum_{Y_{i+1}\in\{0,1\}} p(Y_{i+1}|\mathbf{y}^i)\log_2\frac{1}{p(Y_{i+1}|\mathbf{y}^i)} \tag{4.6}$$

$$H(Y_{i+1}|\mathbf{w},\mathbf{y}^i) = \sum_{Y_{i+1}\in\{0,1\}} p(Y_{i+1}|\mathbf{w})\log_2\frac{1}{p(Y_{i+1}|\mathbf{w})} \tag{4.7}$$

$$H(Y|\mathbf{W},\mathbf{y}^i) = \mathbb{E}_{\mathbf{W}|\mathbf{y}^i}[H(Y_{i+1}|\mathbf{W}=\mathbf{w},\mathbf{y}^i)]. \tag{4.8}$$

Unlike the MSE and posterior volume strategies, mutual information estimation only requires a *single* batch of posterior samples at each round of query selection, which is used to estimate the discrete entropy quantities in (4.6)–(4.8). If S samples are drawn from the posterior distribution, the information in (4.5) can be estimated for a single pair in $O(dS)$

operations. For a pool of M candidate pairwise queries pairs, a pair can be selected in $O(dSM)$ operations at each iteration, which scales linearly in the number of samples used to estimate mutual information. Even though this is still more computationally feasible than direct MSE or posterior volume minimization, a computational complexity of $O(dSM)$ results in an expensive number of computations for highly accurate mutual information estimates over a large pool of candidate pairs.

Because of these analytical and computational challenges with this direct mutual information maximization strategy, we develop two heuristics that mimic the action of maximizing mutual information. In the next section we describe our first heuristic, which we analyze for more refined upper and lower bounds on the number of queries needed to shrink the posterior to a desired volume. Then we introduce a second heuristic which achieves a reduced computational complexity while still remaining theoretically coupled to mutual information maximization.

4.3.2 Heuristic 1: equiprobable, max-variance

In developing a heuristic for mutual information maximization, consider the scenario where *arbitrary* pairs can be generated (unconstrained to a given dataset), resulting in a bisecting hyperplane parameterized by weights a and threshold b. In practice, such queries correspond to the generation of synthetic items in the low-dimensional latent space with tools such as generative adversarial networks [167]. With the freedom to select any hyperplane, consider an *equiprobable* query strategy where where b is selected such that each item in the query will be selected with probability 1/2. This strategy is motivated from the fact that mutual information is upper bounded by $H(Y_{i+1}|\mathbf{y}^i)$, which is maximized at 1 bit if and only if the response probability is equiprobable [166].

To motivate the selection of query hyperplane directions, we define a query's *projected variance* denoted as σ_i^2 as the variance of the posterior marginal in the direction of a query's hyperplane, i.e. $\sigma_i^2 = \sqrt{\mathbf{a}_i^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i} \mathbf{a}_i}$. This corresponds to a measure of how far away the user point is from the hyperplane query, in expectation over the posterior distribution. With

this notation, we have the following lower bound on mutual information for equiprobable queries.

Proposition 4.3.3. *For any "equiprobable" query scheme, for any choice of* $0 \leq c \leq 1$ *we have*

$$I(\mathbf{W}; Y_i|\mathbf{y}^{i-1}) \geq \left(1 - h_b\left(f\left(\frac{ck\sigma_i}{2}\right)\right)\right)(1-c) =: L_{c,k}(\sigma_i)$$

where $\sigma_i = \sqrt{\mathbf{a}_i^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i} \mathbf{a}_i}$ *and* $h_b(p) = p\log_2(1/p) + (1-p)\log_2(1/(1-p))$ *is the binary entropy function.*

This lower bound is monotonically increasing with $k\sigma_i$ and achieves maximum mutual information of 1 bit at $k \to \infty$ and/or $\sigma_i \to \infty$ (with an appropriate choice of c). This suggests choosing weights a that *maximize projected variance* in addition to selecting b according to the equiprobable strategy. Together, we refer to the selection of equiprobable queries in the direction of largest posterior variance as the equiprobable-max-variance, or EPMV scheme for short.

Our primary result concerns the expected number of comparisons (or query complexity) sufficient to reduce the posterior volume below a specified threshold set a priori, using EPMV.

Theorem 4.3.4. *For the EPMV query scheme with* $k\|\mathbf{a}_i\| \geq k_{min} > 0$ *for each selected query with hyperplane weights* $\mathbf{a}_i$*, consider the stopping time* $T_\varepsilon = \min\{i : |\boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}|^{\frac{1}{d}} < \varepsilon\}$ *for stopping threshold* $\varepsilon > 0$*. For* $\tau_1 = \frac{d}{2}\log_2(\frac{1}{2\pi e\varepsilon})$ *and* $\tau_2 = \frac{d}{2}\log_2\frac{c_d}{2\varepsilon} + d$*, we have*

$$\tau_1 \leq E[T_\varepsilon] \leq \tau_2 + \frac{\tau_2 + 1}{l(\tau_2)} - \frac{1}{l(\tau_2)}\int_0^{\tau_2} l(x)dx$$

where $l(x) = L_{c,k_{min}}\left(\frac{2^{\frac{-x}{d}}}{\sqrt{2\pi e}}\right)$ *for any* $0 \leq c \leq 1$ *as defined in Proposition 4.3.3. Furthermore, the lower bound is true for any query selection scheme.*

This result follows from a martingale stopping-time analysis of the entropy at each stage. Our next theorem is less general, but is more easily interpretable.

Theorem 4.3.5. *The EPMV scheme, under the same assumptions as in Theorem 4.3.4, satisfies*

$$\mathbb{E}[T_\varepsilon] = O\left(d \log \frac{1}{\varepsilon} + \left(\frac{1}{\varepsilon k_{min}^2}\right) d \log \frac{1}{\varepsilon}\right).$$

Furthermore, for any query scheme,

$$\mathbb{E}[T_\varepsilon] = \Omega\left(d \log \frac{1}{\varepsilon}\right).$$

This result has a favorable dependence on dimension d, and can be interpreted as a blend between two rates, one of which matches that of the generic lower bound. Our upper bound provides some evidence that our ability to recover w worsens considerably as $k_{\min}$ decreases as the second term is dominant. This is intuitively unsurprising since small $k_{\min}$ corresponds to the case where queries are very noisy. In the derivation of the lower bound, a naïve upper bound of 1 bit is used on mutual information. In actuality, we expect a generic upper bound on mutual information that *decreases* with small $k_{\min}$, due to a resulting increase in response noise. Thus, we expect a similar penalty term in the complexity of the generic lower bound, which is deferred to future work.

On the other hand, for asymptotically large k, we have the following corollary.

Corollary 4.3.6. *In the noiseless setting ($k_0 \to \infty$), EPMV has optimal expected stopping time complexity for posterior volume stopping.*

Proof. For $k_0 \to \infty$, by definition $k_{\min} \to \infty$ (ignoring the pathological case of a single item being compared to itself i.e. $\|a\| = 0$). In this case from Theorem 4.3.5, $\mathbb{E}[T_\varepsilon] = O\left(d \log \frac{1}{\varepsilon}\right)$, which is optimal since for any scheme, $\mathbb{E}[T_\varepsilon] = \Omega\left(d \log \frac{1}{\varepsilon}\right)$. □

Taken together, these bounds suggest that EPMV is optimal up to a penalty term which decreases to zero for large noise constants.

While EPMV was derived under the assumption of arbitrary hyperplane queries, depending on the application we may have to select a pair from a fixed pool of item combinations in a given dataset. For this purpose, we implement a metric for any given pair that, when

maximized over all pairs in a pool, approximates the behavior of EPMV. For a pair with items $\mathbf{p}$ and $\mathbf{q}$ in item pool $\mathscr{P}$, let $\mathbf{a}_{pq} = 2(\mathbf{p} - \mathbf{q})$ and $b_{pq} = \|p\|^2 - \|q\|^2$ denote the weights and threshold parameterizing the bisecting hyperplane. We then implement EPMV with the heuristic function (for some $\lambda > 0$):

$$\eta(\mathbf{p}, \mathbf{q}; \widetilde{\mathbf{W}}) = k_{pq} \sqrt{\mathbf{a}_{pq}^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{Y}^i} \mathbf{a}_{pq}} - \lambda \left| \hat{p}_1 - \frac{1}{2} \right|. \tag{4.9}$$

where

$$\hat{p}_1 = \mathbb{E}\, f(k_{pq}(\mathbf{a}_{pq}^T \mathbf{W} - b_{pq})).$$

This has the effect of selecting queries which are close to equiprobable and align with the direction of largest variance, weighted by k_{pq} to prefer higher fidelity queries. However, $\hat{p}_1$ must be estimated for each pair, requiring $O(dS)$ operations per pair and $O(dSM)$ operations in total, which is the same computational complexity as the naïve information maximization approach. For this reason, we develop a second heuristic that approximates EPMV while significantly reducing the computational complexity.

4.3.3 Heuristic 2: mean-cut, max-variance

Our second heuristic is a *meancut* strategy where b is selected such that the hyperplane passes through the mean of $p(\mathbf{W}|\mathbf{Y}^i)$, i.e. $\mathbf{a}^T\, \mathbb{E}[\mathbf{W}|\mathbf{Y}^i] - b = 0$. For such a strategy, we have the following proposition:

Proposition 4.3.7. *For meancut queries parameterized by hyperplane weight a and noise constant k we have*

$$\left| p(Y_{i+1}|\mathbf{y}^i) - \frac{1}{2} \right| \leq \frac{e-2}{2e} + \frac{\ln 2}{k\sqrt{\mathbf{a}^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i} a}}$$

and

$$I(\mathbf{W}; Y|\mathbf{y}^i) \geq h_b\left(\frac{1}{e} - \frac{\ln 2}{k\sqrt{\mathbf{a}^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i} a}} \right) - \frac{\pi^2 (\log_2 e)}{3k\sqrt{\mathbf{a}^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i} a}}$$

For large projected variances, we observe that $|p(Y_{i+1}|\mathbf{y}^i) - \frac{1}{2}| \lessapprox 0.14$, suggesting that meancut queries are somewhat of an approximation to equiprobable queries in this setting.

Furthermore, notice that the lower bound to mutual information in Proposition 4.3.7 is a monotonically increasing function of the projected variance, $\mathbf{a}^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i} a$. As $\mathbf{a}^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i} a \to \infty$, this bound approaches $h_b(1/e) \approx 0.95$ which is nearly sharp since a query's mutual information is upper bounded by 1 bit. This implies some correspondence between maximizing a query's mutual information and maximizing the projected variance, as was the case in EPMV. Hence, this combination of *meancut, maximum variance* queries (referred to as MCMV) serves as an approximation to EPMV while still maximizing a lower bound on mutual information.

For implementing MCMV over a fixed pool of pairs (rather than arbitrary hyperplanes), we calculate the orthogonal distance of each pair's hyperplane to posterior mean as $|\mathbf{a}_{pq}^T \mathbb{E}[\mathbf{W}|\mathbf{Y}^i] - b_{pq}|/\|\mathbf{a}_{pq}\|_2$, and the projected variance as $\mathbf{a}_{pq}^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{Y}^i} \mathbf{a}_{pq}$. Since it is unlikely that any given pair simultaneously minimizes the hyperplane distance to the posterior mean, has a maximal noise constant, and maximizes projected variance, each pair is evaluated with respect to a tradeoff between these parameters (for some parameter $\lambda > 0$):

$$\eta(p, q; \mu, \boldsymbol{\Sigma}) = k_{pq} \sqrt{\mathbf{a}_{pq}^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{Y}^i} \mathbf{a}_{pq}} - \lambda \frac{|\mathbf{a}_{pq}^T \mathbb{E}[\mathbf{W}|\mathbf{Y}^i] - b_{pq}|}{\|\mathbf{a}_{pq}\|_2} \tag{4.10}$$

This strategy seeks pairs that have large projected variances, while still cutting close to the mean of the posterior. This heuristic is attractive from a computational standpoint since the posterior mean $\mathbb{E}[\mathbf{W}|\mathbf{Y}^i]$ and covariance $\boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{Y}^i}$ can be estimated *once* in $O(d^2 S)$ computations, and then calculating the hyperplane distance from mean and projected variance $\mathbf{a}^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{Y}^i} a$ requires only $O(d^2)$ computations per pair. Overall, this implementation of the MCMV strategy has a computational complexity of $O(d^2(S + M))$, which scales more favorably than both the naïve mutual information estimation and EPMV strategies.

We unify the naïve, EPMV, and MCMV query selection methods into a single framework described in Algorithm 1. At each round of querying, a pair is selected that maximizes an information metric $\eta(\mathbf{p}, \mathbf{q})$ over a pool of candidate pairs. For naïve mutual information

maximization, $\eta(\mathbf{p}, \mathbf{q}) = I(\mathbf{W}; Y_{i+1}|\mathbf{y}^i)$. In Algorithm 1 we use $\eta(\mathbf{p}, \mathbf{q}; \widetilde{\mathbf{W}}) \approx \eta(p, q)$ where $\widehat{\mathbf{W}}$ is a batch of posterior samples. (4.9) and (4.10) define $\eta(\mathbf{p}, \mathbf{q})$ for EPMV and MCMV, respectively.

Algorithm 1 User preference search with noisy paired comparisons

Input: item set $\mathscr{X}$
  $\mathscr{P} \leftarrow$ set of all pairwise queries from items in $\mathscr{X}$
  $\mathscr{P}_\beta \leftarrow$ uniformly downsample $\mathscr{P}$ at rate $0 < \beta \leq 1$
  $\mathbf{y}^0 \leftarrow \varnothing$ initialize set of user responses
  for i = 1 **to** T **do**
    $\widetilde{\mathbf{W}}_i \leftarrow$ batch of S samples from posterior $\mathbf{W}|\mathbf{Y}^{i-1}$
    $\boldsymbol{\mu}_i \leftarrow \mathbb{E}[\mathbf{W}|\mathbf{Y}^{i-1}]$
    $\boldsymbol{\Sigma}_i \leftarrow \mathbb{E}[(\mathbf{W} - \boldsymbol{\mu}_i)(\mathbf{W} - \boldsymbol{\mu}_i)^T|\mathbf{Y}^{i-1}]$
    if Naïve method **then**
      $\mathbf{p}_i, \mathbf{q}_i \leftarrow \arg\max_{p,q \in \mathscr{P}_\beta} \eta(\mathbf{p}, \mathbf{q}; \widetilde{\mathbf{W}}_i)$
    else if MPMV method **then**
      $\mathbf{p}_i, \mathbf{q}_i \leftarrow \arg\max_{p,q \in \mathscr{P}_\beta} \eta(\mathbf{p}, \mathbf{q}; \boldsymbol{\Sigma}_i, \widetilde{\mathbf{W}}_i)$
    else if MCMV method **then**
      $\mathbf{p}_i, \mathbf{q}_i \leftarrow \arg\max_{p,q \in \mathscr{P}_\beta} \eta(\mathbf{p}, \mathbf{q}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
    end if
    $\mathbf{y}_i \leftarrow$ PairedComparison$(\mathbf{p}_i, \mathbf{q}_i)$
    $\mathbf{y}^i \leftarrow [y_i, \mathbf{y}^{i-1}]$
    $\widehat{\mathbf{W}}_i \leftarrow \mathbb{E}[\mathbf{W}|\mathbf{Y}^i]$
  end for
Output: $\widehat{\mathbf{W}}_T$

4.4 Simulation results

To evaluate our approach, we constructed a realistic embedding (from a set of training user-response triplets) consisting of multi-dimensional item points and simulated our pairwise search methods over randomly generated user points. We constructed an item embedding of the Yummly `Food-10k` data-set of [168, 169], consisting of 958,479 publicly available triplet comparisons assessing relative similarity among 10,000 food items. The item coordinates are derived only from the crowd-sourced triplets using the algorithm of [155]. We note that the embedding method used assumes a slightly different probabilistic model than we do in this chapter. Specifically, the probability that item $\mathbf{r}$ is rated closer to $\mathbf{p}$ than to $\mathbf{q}$

is given by

$$\mathbb{P}'(\mathbf{p} \prec \mathbf{q} \mid \mathbf{r}) = \frac{1}{1 + \|\mathbf{r} - \mathbf{p}\|^2 / \|\mathbf{r} - \mathbf{q}\|^2}.$$

Contrast this with (4.1) which supposes

$$\mathbb{P}(\mathbf{p} \prec \mathbf{q} \mid \mathbf{r}) = \frac{1}{1 + \exp(\|\mathbf{r} - \mathbf{p}\|^2 - \|\mathbf{r} - \mathbf{q}\|^2)}.$$

Despite this difference, the successful performance of our simulation results show that our method is somewhat adaptive to differing methods used for embedding construction.

In each simulation trial, we generate

$$\mathbf{W} \sim \text{Unif}([-1, 1]^d),$$

then adaptively collect paired comparisons drawn from the item points in our embedding, according to Algorithm 1 using the discussed heuristic functions η. As a baseline non-adaptive method, we also perform simulations using a randomly selected pair for each query. The response probability of each observation follows (4.1), using each of the three schemes for choosing k_{pq} described in (K1)–(K3). In each noise model we optimized the value of k_0 using maximum-likelihood estimation (according to our probabilistic model (4.1) over a set of training triplets. We use the Stan Modeling Language [170] to generate posterior samples, since our model is particularly amenable to MCMC methods [171] due to log-concavity of posteriors.

While it would be possible to compare our strategies against the method in [135] since it produces a user point *cell* in $\mathbb{R}^d$ as in Figure 4.1, from which a user preference can be estimated through a statistic such as the cell's centroid. However, the estimation error of this method is fundamentally lower bounded by the diameter of each cell, which is fixed and determined *a priori* by item embedding positions. Since this constraint precludes arbitrary continuous estimation of a user point (which is the goal of our work), we omit a comparison to this method in our simulations.

In Figure 4.2 we plot MSE versus iteration number (i.e. number of queries asked) av-

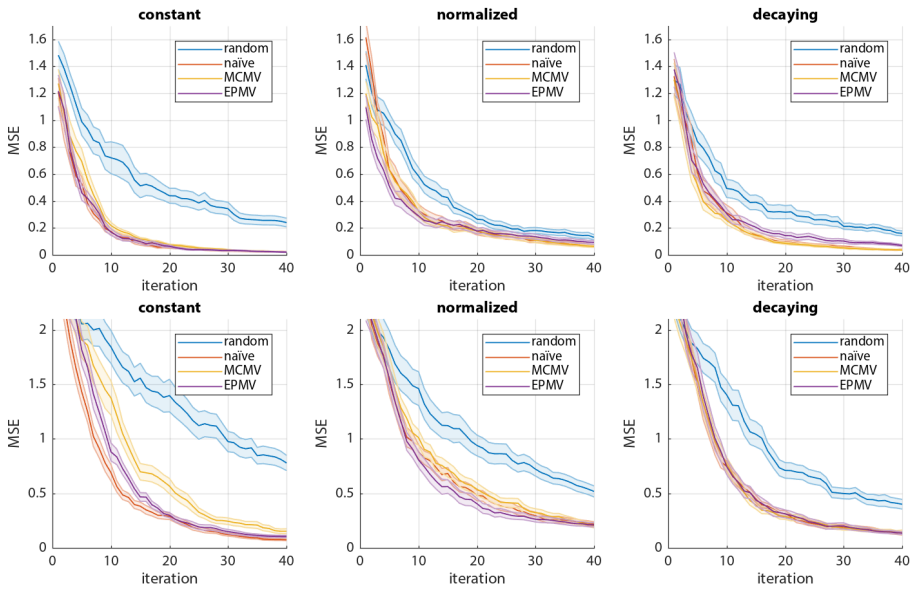


Figure 4.2: Mean squared error $\pm$ one standard error for various query selection methods and noise constant models in a search task of the Yummly `Food-10k` data-set embeddings of dimension 4 (top row) and dimension 7 (bottom row). Note that the three adaptive methods perform similarly in many cases and sometimes overlap.

eraged over 30 trials of each method and each of the k_{pq} assumptions (K1)–(K3) for the Food-10k data-set embedding. In every case, adaptive selection outperforms random selection by a substantial margin. The EPMV and MCMV methods perform similarly to the naïve mutual information maximization method, with a steep decrease in MSE over the initial queries followed by a gradual decrease in decay rate. These trends are evident across all noise constant models, and across embeddings of various dimensions.

In Figure 4.3, we give a qualitative visualization of the error rate of the naïve and random query selection schemes. For each method we employ the 'decaying' noise constant model and select the trial whose final error was closest to each method's respective MSE. In each image, the 20 nearest neighbors to the user point estimate after 40 queries are displayed, along with the item closest to ground truth user point, displayed on top. In the naïve selection case, the closest item to the user is also the closest item to the estimate, suggestion that the method successfully estimates the neighborhood of the true user point. In the random case however, the ground truth's nearest neighboring item is not even in the top 20 closest

items to the estimate, implying that the random query strategy fails to efficiently learn the neighborhood of the user point. This sheds light on the fact that the margin in MSE between random and adaptive methods can lead to substantial differences in the perceptual fidelity of the user estimate point in embedding space.

Figure 4.3: Top-20 neighborhood after 40 queries against user point nearest neighbor for closest-to-average performing trials in random (left) and naïve mutual information maximization (right) query selection schemes. The random method fails to identify a neighborhood that includes the user point's nearest neighbor item, while the naïve adaptive method succeeds.

Furthermore, this margin can lead to substantial differences in the size of the neighborhood of items around a user estimate. For instance, by looking at the ball centered at the embedding origin with the final MSE as its squared-radius, this ball only contains 35 items for the naïve method but contains 216 items for randomly generated queries. This difference is substantial when a user may wish to terminate search after a fixed number of iterations to generate a neighborhood of items and manually select the item most similar to their preference, since a user would need to manually select from a much larger pool of items after random queries than for the naïve information maximization strategy.

4.5 Discussion

Our simulated results demonstrate that both heuristic methods, EPMV and MCMV, greatly outperform the case where pairs are chosen randomly. Interestingly, the MCMV approach, while being extremely simple to implement as well as computationally efficient, performs nearly as well as using a naïve mutual information estimate. On the other hand, as suspected from Proposition 4.3.7, MCMV also performs similarly to EPMV, a method for which we

have provided information theoretic performance guarantees. This motivates further study towards direct performance guarantees for the MCMV heuristic.

Although in this work pairs were drawn from a fixed item embedding, we note that this was not essential for the MCMV method and that it is adaptable to continuous item spaces, which would allow for generative construction of proposal pairs. This is a very reasonable assumption in some applications, such as facial composite generation for criminal cases [172] or in evaluating foods and beverages, where we have the ability to generate almost arbitrary stimuli based on the ratios of ingredients [173].

4.6 Proof details

4.6.1 Proof of Lemma 4.3.1

First, we begin with an additional lemma:

Lemma 4.6.1. *Let X_i be a marginal distribution of* $\mathbf{W}$. *The density of X_i is then*

$$p_{X_i|\mathbf{y}^{i-1}}(x) = \frac{1}{\sigma_i} p_{Z_i}\left(\frac{X_i - \mathbb{E}[X_i|\mathbf{y}^i]}{\sigma_i}\right) \leq \frac{1}{\sigma_i},$$

where $\sigma_i = \sqrt{\mathbb{E}[(X_i - \mathbb{E}[X_i|\mathbf{y}^{i-1}])^2|\mathbf{y}^i]}$ *and* $Z_i = \frac{X_i - \mathbb{E}[X_i|\mathbf{y}^i]}{\sigma_i}$.

Proof. Since X_i is a marginal of a log-concave distribution, X_i is also log-concave. Furthermore, Z_i is a zero-mean, unit-variance (i.e., isotropic) log-concave random variable with density $p_{Z_i}(z)$. Then Lemma 4.6.1 follows because one-dimensional isotropic log-concave densities are upper bounded by one [174]. □

A direct consequence of Lemma 4.6.1 is that for any $a > 0$,

$$\begin{aligned}\mathbb{P}(|X_i| < a \,|\, \mathbf{y}^{i-1} &= \int_{-a}^{a} p_{X_i|\mathbf{y}^{i-1}}(x)dx \\ &\leq \frac{1}{\sigma_i}\int_{-a}^{a} dx \leq \frac{2a}{\sigma_i}\end{aligned}$$

implying that

$$\mathbb{P}(|X_i| \geq a \,|\, \mathbf{y}^{i-1} \geq 1 - \frac{2a}{\sigma_i}. \tag{4.11}$$

Proof of Lemma 4.3.1. Letting $\mathbf{\Sigma_W}$ denote the $d \times d$ covariance matrix of random vector $\mathbf{W} \in \mathbb{R}^d$, from Theorem 8.6.5 in [166], we have the upper bound

$$h(\mathbf{W}) \leq \frac{1}{2} \log_2((2\pi e)^d |\mathbf{\Sigma_W}|). \tag{4.12}$$

Now assume the distribution $P_\mathbf{W}$ of $\mathbf{W}$ is log-concave, let $W_1, W_2 \sim P_\mathbf{W}$ be i.i.d. and let $\widetilde{W} := W_1 - W_2$. Let $p_{\widetilde{\mathbf{W}}}$ and $p_\mathbf{W}$ denote the respective densities of $\widetilde{\mathbf{W}}$ and $\mathbf{W}$. We have by Proposition 3.5 of [165], for all $\mathbf{z} \in \mathbb{R}^d$,

$$p_{\widetilde{\mathbf{W}}}(\mathbf{z}) = p_\mathbf{W}(\mathbf{z}) \star p_\mathbf{W}(-\mathbf{z}), \tag{4.13}$$

where $\star$ is the convolution operator, is also log-concave. Since covariances add for independent random vectors, $\mathbf{\Sigma}_{\widetilde{\mathbf{W}}} = 2\mathbf{\Sigma_W}$.

By Theorem 4 of [175],

$$h(\widetilde{\mathbf{W}}) \geq \frac{d}{2} \log_2 \frac{|\mathbf{\Sigma}_{\widetilde{\mathbf{W}}}|^{1/d}}{c(d)},$$

where $c(d) = e^2 d^2/(4\sqrt{2}(d+2))$. From Corollary 2.3 of [176],

$$h(\widetilde{\mathbf{W}}) = h(\mathbf{W}_1 - \mathbf{W}_2) \leq h(\mathbf{W}) + d,$$

which implies

$$\begin{aligned} h(\mathbf{W}) \geq h(\widetilde{\mathbf{W}}) - d &\geq \frac{d}{2} \log_2 \frac{|\mathbf{\Sigma}_{\widetilde{\mathbf{W}}}|^{1/d}}{c(d)} - d \\ &\geq \frac{d}{2} \log_2 \frac{|2\mathbf{\Sigma_W}|^{1/d}}{c(d)} - d \end{aligned} \tag{4.14}$$

The result follows combining (4.13) and (4.14). □

4.6.2 Proof of Theorem 4.3.2

$$\mathbb{E}_{\mathbf{Y}^i}[h_i(\mathbf{W})] = h_0(\mathbf{W}) - \sum_{j=1}^{i} I(\mathbf{W}; Y_j | Y^{j-1}) \tag{4.15}$$

$$\geq -i \tag{4.16}$$

from the chain rule for mutual information with $h_0(\mathbf{W}) = 0$ and $I(\mathbf{W}; Y_j|Y^{j-1}) \leq 1$ [166], and

$$\mathbb{E}_{\mathbf{Y}^i}[h_i(\mathbf{W})] \leq \frac{1}{2}\,\mathbb{E}_{\mathbf{Y}^i} \log_2((2\pi e)^d |\mathbf{\Sigma}_{\mathbf{W}|\mathbf{Y}^i}|) \tag{4.17}$$

$$\leq \frac{1}{2} \log_2((2\pi e)^d |\mathbb{E}_{\mathbf{Y}^i}\, \mathbf{\Sigma}_{\mathbf{W}|\mathbf{Y}^i}|) \tag{4.18}$$

from Lemma (4.3.1) with Jensen's inequality and the concavity of $\log|A|$ for any matrix A in the positive definite cone [105]. Rearranging, we have

$$\frac{2^{-2i}}{(2\pi e)^d} \leq |\mathbb{E}_{\mathbf{Y}^i}\, \mathbf{\Sigma}_{\mathbf{W}|\mathbf{Y}^i}| \tag{4.19}$$

$$\leq \frac{\left(\mathrm{Tr}\left(\mathbb{E}_{\mathbf{Y}^i}[\mathbf{\Sigma}_{\mathbf{W}|\mathbf{y}^i}]\right)^d}{d^d} \tag{4.20}$$

$$= \frac{(\mathbb{E}_{\mathbf{W},\mathbf{Y}^i}[\|\mathbf{W} - \mathbb{E}[\mathbf{W}|\mathbf{Y}^i]\|_2^2])^d}{d^d} \tag{4.21}$$

$$\leq \frac{(\mathbb{E}_{\mathbf{W},\mathbf{Y}^i}[\|\mathbf{W} - \widehat{\mathbf{W}}_i^2]\|)^d}{d^d} \tag{4.22}$$

where (4.20) is from the AM–GM inequality, (4.21) is due to the linearity of trace and expectation, and the last inequality is due to that fact that expected value is the MMSE estimator, from which the MSE lower bound follows.

4.6.3 Proof of Proposition 4.3.3

Proof. Consider the 'equiprobable' query scheme, with $\mathbb{P}(Y_i = 1|\mathbf{y}^{i-1}) = \frac{1}{2}$ for hyperplane query given by weights $\mathbf{a}_i$, threshold τ_i, and noise constant k. Letting $X_i = \mathbf{a}_i^T W - \tau_i$, we have

$$\begin{aligned} I(\mathbf{W}; Y_i|\mathbf{y}^{i-1}) &= H(Y_i|\mathbf{y}^{i-1}) - H(Y_i|\mathbf{y}^{i-1}, W) \\ &= H(Y_i|\mathbf{y}^{i-1}) - H(Y_i|\mathbf{y}^{i-1}, W, X_i) \end{aligned}$$

since X_i is a deterministic function of W

$$= H(Y_i|\mathbf{y}^{i-1}) - H(Y_i|\mathbf{y}^{i-1}, X_i)$$

since $p(Y_i|\mathbf{y}^{i-1}, W, X_i) = p(Y_i|\mathbf{y}^{i-1}, X_i)$

$$= I(X_i; Y_i|\mathbf{y}^{i-1}).$$

Revisiting mutual information, we have

$$I(X_i; Y_i|\mathbf{y}^{i-1}) = \mathbb{E}\left[\log_2 \frac{p(Y_i|X_i, \mathbf{y}^{i-1})}{p(Y_i)}\right] \tag{4.23}$$

$$= E_{X_i}[(1 - h_b(f(kX_i))) \,|\, \mathbf{y}^{i-1}] \tag{4.24}$$

$$= E_{X_i}[(1 - h_b(f(k|X_i|)))|\mathbf{y}^{i-1}] \tag{4.25}$$

since $1 - h_b(f(kX_i))$ is symmetric. From Markov's inequality with $1 - h_b(f(k|X_i|))$ being monotonically increasing, for any $a > 0$,

$$\geq (1 - h_b(f(ka)))\, \mathbb{P}(|X| > a \,|\, \mathbf{y}^{i-1}) \tag{4.26}$$

$$\text{from (4.11))} \quad \geq (1 - h_b(f(ka)))\left(1 - \frac{2a}{\sigma_i}\right) \tag{4.27}$$

$$= \left(1 - h_b\left(f\left(\frac{kc\sigma_i}{2}\right)\right)\right)(1 - c) \tag{4.28}$$

by letting $a = \frac{c\sigma_i}{2}$ for any $0 \leq c \leq 1$ □

4.6.4 Proof of Theorem 4.3.4

Entropy Properties:.— Let $h(\mathbf{W}|\mathbf{y}^i)$ denote the posterior entropy after observing i queries. With a uniform prior distribution over the hypercube $[-\frac{1}{2}, \frac{1}{2}]$, we have that $h(\mathbf{W}|\mathbf{y}^0) = 0$ and $h(\mathbf{W}|\mathbf{y}^i) \leq 0$ for $\forall i$ since the uniform distribution maximizes entropy over this bounded space.

After query i, let the eigenvalues of the posterior covariance matrix be denoted in decreasing order as $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_d$. In the equiprobable, max-variance scheme, query $\mathbf{a}_i$ is in the direction of maximal eigenvector, so the product of the noise constant and query standard deviation at iteration i is given by $k\sqrt{\mathbf{a}_i^T \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i} \mathbf{a}_i} = k\|\mathbf{a}_i\|\sqrt{\lambda_1} \geq k_{\min}\sqrt{\lambda_1}$. From

the monotonicity of the mutual information lower bound on equiprobable queries, we have

$$I(\mathbf{W}; Y_i|\mathbf{y}^{i-1}) \geq L_{c,k_{\min}}(\sqrt{\lambda_1}) \tag{4.29}$$

From rearranging terms in Lemma 4.3.1 along with $|\mathbf{\Sigma}_{\mathbf{W}|\mathbf{y}^i}| = \prod_{i=1}^{d} \lambda_i$, we have

$$\frac{2^{2h(\mathbf{W}|\mathbf{y}^i)}}{(2\pi e)^d} \leq |\mathbf{\Sigma}_{\mathbf{W}|\mathbf{y}^i}| = \prod_{i=1}^{d} \lambda_i \leq \lambda_1^d \tag{4.30}$$

$$\implies \lambda_1 \geq \frac{2^{\frac{2h(\mathbf{W}|\mathbf{y}^i)}{d}}}{2\pi e} \tag{4.31}$$

For compactness of notation, let

$$\tilde{L}_{c,k_{\min}}(h) = L_{c,k_{\min}}\left(\frac{2^{\frac{h}{d}}}{\sqrt{2\pi e}}\right) \tag{4.32}$$

Since $L_{c,k_{\min}}$ is monotonically increasing, we have

$$I(\mathbf{W}; Y_i|\mathbf{y}^{i-1}) \geq \tilde{L}_{c,k_{\min}}(h(\mathbf{W}|\mathbf{y}^i)) \tag{4.33}$$

Combined with the 1 bit upper bound on mutual information along $I(\mathbf{W}; Y_i|\mathbf{y}^{i-1}) = h(\mathbf{W}|\mathbf{y}^i) - \mathbb{E}_{Y_{i+1}|\mathbf{y}^i}[h(\mathbf{W}|\mathbf{y}^{i+1})]$, we have

$$\begin{aligned} h(\mathbf{W}|\mathbf{y}^i) - 1 &\leq \mathbb{E}_{Y_{i+1}|\mathbf{y}^i}[h(\mathbf{W}|\mathbf{y}^{i+1})] \\ &\leq h(\mathbf{W}|\mathbf{y}^i) - \tilde{L}_{c,k_{\min}}(h(\mathbf{W}|\mathbf{y}^i)) \end{aligned} \tag{4.34}$$

To bound the entropy deviations from one measurement to the next, we need the following lemma:

Lemma 4.6.2. *For the equiprobable query scheme,*

$$|h(\mathbf{W}|\mathbf{y}^{i+1}) - h(\mathbf{W}|\mathbf{y}^i)| \leq \gamma(d) \quad \forall i \geq 0$$

where $\gamma(d) = 8d + \frac{d}{2}\log_2(2\pi e d) + 1$.

The proof of Lemma 4.6.2 is highly technical and so we relegate it to the end of the

supplementary materials.

Martingale Properties:.— Let $Z_i = -h(\mathbf{W}|\mathbf{y}^i)$. From the previous section we have $Z_0 = 0$, $Z_i \geq 0 \; \forall i \geq 0$, $|Z_{i+1} - Z_i| \leq \gamma(d)$ from Lemma 4.6.2, and $Z_i + \tilde{L}_{c,k_{\min}}(-Z_i) \leq E_{Z_{i+1}|\mathbf{y}^i}[Z_{i+1}] \leq Z_i + 1$. Since Z_i is a deterministic function of $\mathbf{y}^i \; \forall i$ along with the law of total expectation,

$$\begin{aligned}\mathbb{E}[Z_{i+1}|Z_0, \dots, Z_i] &= \mathbb{E}_{\mathbf{Y}^i|Z_0,\dots,Z_i}\, \mathbb{E}[Z_{i+1}|Z_0, \dots, Z_i, \mathbf{y}^i] \\ &= \mathbb{E}_{\mathbf{Y}^i|Z_0,\dots,Z_i}\, \mathbb{E}[Z_{i+1}|\mathbf{y}^i]\end{aligned}$$

which implies

$$\begin{aligned}\mathbb{E}[Z_{i+1}|Z^i] &\geq \mathbb{E}_{\mathbf{Y}^i|Z_0,\dots,Z_i}[Z_i + \tilde{L}_{c,k_{\min}}(-Z_i)] \\ &= Z_i + \tilde{L}_{c,k_{\min}}(-Z_i)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[Z_{i+1}|Z_0, \dots, Z_i] &\leq \mathbb{E}_{\mathbf{Y}^i|Z_0,\dots,Z_i}[Z_i + 1] \\ &= Z_i + 1\end{aligned}$$

Since $\tilde{L}_{c,k_{\min}}(-Z_i) > 0$, we have $\mathbb{E}[Z_{i+1}|Z^i] \geq Z_i$. For all $i \geq 0$, $|Z_i| < \infty$ since $|Z_i| = |Z_0 + \sum_{j=1}^{i} Z_j - Z_{j-1}| \leq \sum_{j=1}^{i} |Z_j - Z_{j-1}| \leq i\gamma(d) < \infty$. Therefore, Z_i is a submartingale.

Let $\tau > 0$ define a stopping threshold and corresponding stopping time $T = \min\{i : Z_i \geq \tau\}$ Considering $E[Z_{i+1}|Z^i] \leq Z_i + 1$ and taking the expectation over Z^i on both

sides and expanding, we have

$$\begin{aligned}
\mathbb{E}[E[Z_{i+1}|U^i]] &\leq \mathbb{E}[Z_i] + 1 \\
\mathbb{E}[Z_{i+1}] &\leq \mathbb{E}\,\mathbb{E}[Z_i|Z^{i-1}] + 1 \\
\mathbb{E}[Z_{i+1}] &\leq \mathbb{E}[Z_{i-1}] + 1 + 1 \\
&\cdots \\
\mathbb{E}[Z_{i+1}] &\leq i + 1 \\
\implies \mathbb{E}[Z_i] &\leq i
\end{aligned}$$

which implies

$$T \geq \mathbb{E}[Z_T] \geq \tau$$

where the last inequality follows by definition, so $\mathbb{E}[T] \geq \tau$. Note that this is true for *any* query selection scheme since mutual information is always upper bounded by 1 bit.

To lower bound the expected stopping time, observe $\tilde{L}_{c,k_{\min}}(-z)$ is monotonically decreasing in z, and $Z_i \leq \tau$ for $i < T$, we have in this range that $\tilde{L}_{c,k_{\min}}(-Z_i) > \tilde{L}_{c,k_{\min}}(-\tau)$. Using this fact, we construct a separate submartingale that equals Z_i up to and including the stopping time and has the same properties listed above. Specifically, let

$$U_i = \begin{cases} Z_i & i \leq T \\ U_{i-1} + \tilde{L}_{c,k_{\min}}(-\tau) & i > T. \end{cases} \tag{4.35}$$

Clearly for $i \leq T$, $U_i = Z_i$, and if T_U is defined as $T_U = \min\{i : U_i \geq \tau\}$, by observation $T_U = T$. U_i also satisfies $|U_{i+1} - U_i| < \gamma(d)$, and $U_i + \tilde{L}_{c,k_{\min}}(-\tau) \leq E[U_{i+1}|U^i] \leq U_i + 1$.

We have

$$E[U_{i+1}|U^i] \geq U_i + \tilde{L}_{c,k_{\min}}(-\tau) \tag{4.36}$$

$$\frac{E[U_{i+1}|U^i]}{\tilde{L}_{c,k_{\min}}(-\tau)} \geq \frac{U_i}{\tilde{L}_{c,k_{\min}}(-\tau)} + 1 \tag{4.37}$$

$$\frac{E[U_{i+1}|U^i]}{\tilde{L}_{c,k_{\min}}(-\tau)} - (i+1) \geq \frac{U_i}{\tilde{L}_{c,k_{\min}}(-\tau)} - i \tag{4.38}$$

We then have submartingle given by $U_i^{(\text{sub})} = \frac{U_i}{\tilde{L}_{c,k_{\min}}(-\tau)} - i$.

Assume for the time being that the optional stopping theorem can be applied to this submartingale (proved in the sequel) - for any stopping time S satisfying $S \leq T$, $\mathbb{E}[U_S^{\text{sub}}] \leq \mathbb{E}[U_T^{\text{sub}}]$. Specifically, if τ_S is a stopping threshold satisfying $\tau_S \leq \tau$ such that $S = \min\{i : U_i \geq \tau_S\}$, then (for brevity, letting $l(u) = \tilde{L}_{c,k_{\min}}(-u)$)

$$\frac{\mathbb{E}[U_S]}{l(\tau)} - \mathbb{E}[S] \leq \frac{\mathbb{E}[U_T]}{l(\tau)} - \mathbb{E}[S] \tag{4.39}$$

which implies

$$\begin{aligned} \frac{\mathbb{E}[U_S]}{l(\tau_S)} - \mathbb{E}[S] = \frac{l(\tau)}{l(\tau_S)} \left[\frac{\mathbb{E}[U_S]}{l(\tau)} - \mathbb{E}[S] \right] - \\ \left(1 - \frac{l(\tau)}{l(\tau_S)}\right) \mathbb{E}[S] \end{aligned} \tag{4.40}$$

$$\leq \frac{l(\tau)}{l(\tau_S)} \left[\frac{\mathbb{E}[U_T]}{l(\tau)} - \mathbb{E}[T] \right] - \left(1 - \frac{l(\tau)}{l(\tau_S)}\right) \mathbb{E}[S] \tag{4.41}$$

More generally, let $\Delta > 0$ be given and set stopping threshold $\tau_i = i\Delta$, with corresponding stopping time T_i. Define $P_i = \frac{U_{T_i}}{l(\tau_i)} - T_i$. Letting $r_i = \frac{l(\tau_i)}{l(\tau_{i-1})}$ and letting $T = T_i$ and $S = T_{i-1}$, by rearranging the above we have

$$\mathbb{E}[P_i] \geq \frac{\mathbb{E}[P_{i-1}]}{r_i} + \frac{\left(1 - r_i\right)}{r_i} \mathbb{E}[T_{i-1}] \tag{4.42}$$

Noting that $\mathbb{E}[T_0] = 0$ since a threshold of τ_0 results in stopping at $T_0 = 0$ and $E[P_0] =$

$\frac{U_{T_0}}{l(\tau_0)} - \mathbb{E}[T_0] = 0$, we continue this bound recursively

$$\begin{aligned}
\mathbb{E}[P_i] &\geq \frac{\mathbb{E}[P_{i-2}]}{r_i r_{i-1}} + \frac{\left(1 - r_{i-1}\right)}{r_i r_{i-1}} \mathbb{E}[T_{i-2}] \\
&\quad + \frac{\left(1 - r_i\right)}{r_i} \mathbb{E}[T_{i-1}] \dots \\
&= \sum_{j=1}^{i-1} \frac{1 - r_{j+1}}{\prod_{k=j+1}^{i} r_k} \mathbb{E}[T_j] \\
&= \sum_{j=1}^{i-1} \frac{l(\tau_j) - l(\tau_{j+1})}{l(\tau_i)} \mathbb{E}[T_j]
\end{aligned}$$

since $\prod_{k=j+1}^{i} r_k = \frac{l(\tau_i)}{l(\tau_{i-1})} \frac{l(\tau_{i-1})}{l(\tau_{i-2})} \dots \frac{l(\tau_{j+1})}{l(\tau_j)} = \frac{l(\tau_i)}{l(\tau_j)}$

$$\begin{aligned}
&= \frac{1}{l(\tau_i)} \sum_{j=1}^{i-1} \frac{l(j\Delta) - l(j\Delta + \Delta)}{\Delta} \Delta \, \mathbb{E}[T_j] \\
&\geq \frac{1}{l(\tau_i)} \sum_{j=1}^{i-1} \frac{l(\tau_j) - l(\tau_j + \Delta)}{\Delta} \tau_j \Delta
\end{aligned}$$

since $\mathbb{E}[T_j] \geq \tau_j = j\Delta$. Now let $\tau > 0$ be given (with corresponding stopping time T and let $\Delta \to 0$, choosing i appropriately such that $\tau = \tau_i = i\Delta$

$$\begin{aligned}
&\geq -\frac{1}{l(\tau)} \int_0^\tau \frac{d}{dx} l(x) x dx \\
&= \frac{1}{l(\tau)} \int_0^\tau l(x) dx - \tau \\
&\implies \frac{\mathbb{E}[U_T]}{l(\tau)} - \mathbb{E}[T] \geq \frac{1}{l(\tau)} \int_0^\tau l(x) dx - \tau \\
\implies \mathbb{E}[T] &\leq \tau + \frac{\mathbb{E}[U_T]}{l(\tau)} - \frac{1}{l(\tau)} \int_0^\tau l(x) dx \\
&\leq \tau + \frac{\tau + 1}{l(\tau)} - \frac{1}{l(\tau)} \int_0^\tau l(x) dx
\end{aligned}$$

since $\mathbb{E}[U_T] = \mathbb{E}[\mathbb{E}[U_T | U^{T-1}]] \leq \mathbb{E}[U_{T-1}] + 1 \leq \tau + 1$

All together we have

$$\tau \leq \mathbb{E}[T] \leq \tau + \frac{\tau + 1}{l(\tau)} - \frac{1}{l(\tau)} \int_0^{\tau} l(x) dx \tag{4.43}$$

Now, suppose we'd like to stop the algorithm when the posterior covariance determinant crosses below a threshold, corresponding to a low posterior volume. Denote this threshold as ε, and define the stopping time T_ε as $\min\{i : |\mathbf{\Sigma}_{\mathbf{W}|\mathbf{y}^i}|^{\frac{1}{d}} < \varepsilon\}$. By rearranging the upper bound in Lemma 4.3.1 we have the necessary condition

$$h_i(\mathbf{W}) \leq \frac{d}{2} \log_2(2\pi e \varepsilon) \tag{4.44}$$

Letting $\tau_1 = \frac{d}{2} \log_2(\frac{1}{2\pi e \varepsilon})$ be the entropic stopping threshold with stopping time T_1, from (4.43) this results in (with $\mathbb{E}[T_\varepsilon] \geq \mathbb{E}[T_1]$ since this is a necessary condition)

$$\mathbb{E}[T_\varepsilon] \geq \mathbb{E}[T_1] \geq \tau_1 \tag{4.45}$$

Similarly, by rearranging the lower bound in Lemma 4.3.1 we observe that a sufficient condition for this stopping criterion is

$$h_i(\mathbf{W}) \leq \frac{d}{2} \log_2 \frac{2\epsilon}{c_d} - d \tag{4.46}$$

where $c_d = (e^2 d^2)(4\sqrt{2}(d + 2))$. Letting $\tau_2 = \frac{d}{2} \log_2 \frac{c_d}{2\varepsilon} + d$ be the entropic stopping threshold with stopping time T_2, we have from (4.43) (with $\mathbb{E}[T_\varepsilon] \leq \mathbb{E}[T_2]$ since this is only a sufficient condition):

$$\mathbb{E}[T_\varepsilon] \leq \mathbb{E}[T_2] \leq \tau_2 + \frac{\tau_2 + 1}{l(\tau_2)} - \frac{1}{l(\tau_2)} \int_0^{\tau_2} l(x) dx \tag{4.47}$$

Combining these, we have the theorem result.

Verifying Optional Stopping Theorem:.— Consider a submartingale of the form $P_i = \frac{Q_i}{C} - i$ for some $C > 0$, where Q_i is also a submartingale satisfying $Q_i = 0$, $Q_i \geq 0$ for $i \geq 0$,

and $|Q_{i+1} - Q_i| \leq B$ for some $B > C > 0$. This implies

$$\begin{aligned}
|P_{i+1} - P_i| &= \left| \frac{Q_{i+1}}{C} - (i+1) - \frac{Q_i}{C} + i \right| \\
&= \frac{\left|Q_{i+1} - Q_i - C\right|}{C} \\
&\leq \frac{\left|Q_{i+1} - Q_i\right|}{C} + 1 \\
&\leq \frac{B}{C} + 1 =: B' < \infty
\end{aligned}$$

Let stopping time T_Q be defined as $\min\{i : Q_i > \tau\}$ for some threshold $0 < \tau < \infty$. This implies a stopping time on P_i given by $T_P = \min\{i : P_i > \frac{\tau}{C} - i\}$, with $T := T_Q = T_P$. We have from [177, Theorem 5.2.6] that $P_{T\wedge i}$ and $Q_{T\wedge i}$ are also submartingales.

Consider $\sup \mathbb{E}\, Q^+_{T\wedge i} = \sup \mathbb{E}\, Q_{T\wedge i} \leq \tau + B < \infty$, by definition. From [177, Theorem 5.2.8], as $i \to \infty$, $Q_{T\wedge i}$ converges a.s. to Q with $\mathbb{E}[Q] < \infty$.

Similarly, $\sup \mathbb{E}\, P^+_{T\wedge i} = \sup \mathbb{E}\left[\left\{\frac{Q_{T\wedge i}}{C} - (T \wedge i)\right\}^+\right] \leq \sup \mathbb{E}\left[\frac{Q^+_{T\wedge i}}{C}\right] \leq \frac{\tau+B}{C} < \infty$, so as $i \to \infty$, $P_{T\wedge i}$ converges a.s. to P with $\mathbb{E}[P] < \infty$.

We have

$$\begin{aligned}
|T \wedge i| &= \left| (T \wedge i) - \frac{Q_{T\wedge i}}{C} + \frac{Q_{T\wedge i}}{C} \right| \\
&\leq \left| (T \wedge i) - \frac{Q_{T\wedge i}}{C} \right| + \frac{\left|Q_{T\wedge i}\right|}{C} \\
&= |P_{T\wedge i}| + \frac{\left|Q_{T\wedge i}\right|}{C}
\end{aligned}$$

which implies $\lim_{i\to\infty} |T \wedge i| \leq |P| + \frac{|Q|}{C} < \infty$ a.s. This implies that $T < \infty$ a.s., which further implies that $\mathbb{E}[T] < \infty$. Combining this fact with $|P_{i+1} - P_i| \leq B'$, [177, Theorem 5.7.5] gives that $P_{T\wedge n}$ is uniformly integrable. Then, from [177, Theorem 5.7.4], for any stopping time $L \leq T$, $\mathbb{E}[P_L] \leq \mathbb{E}[P_T]$.

4.6.5 Proof of Theorem 4.3.5

To lower bound the complexity of T_ε, we substitute the definition of τ_1 into (4.45), which is true for any query scheme:

$$\mathbb{E}[T_\varepsilon] \geq \frac{d}{2}\log_2\left(\frac{1}{2\pi e\varepsilon}\right) \tag{4.48}$$

$$\implies \mathbb{E}[T_\varepsilon] = \Omega\left(d\log\frac{1}{\varepsilon}\right) \tag{4.49}$$

To upper bound the complexity of T_ε, note that $\tau_2 - \frac{1}{l(\tau_2)}\int_0^{\tau_2} l(x)dx \leq 0$ from the mean value theorem, so $\mathbb{E}[T_\varepsilon] \leq \frac{\tau_2+1}{l(\tau_2)}$. Also note that

$$\begin{aligned} L_{c,k}(\sigma) &= \left(1 - h_b\left(f\left(\frac{ck\sigma}{2}\right)\right)\right)(1-c) \\ &\geq \left(1 - \operatorname{sech}\left(\frac{ck\sigma}{4}\right)\right)(1-c) &(4.50)\\ &\geq \frac{c^2k^2\sigma^2}{32 + c^2k^2\sigma^2}(1-c) &(4.51)\end{aligned}$$

where (4.50) is from $h_b(p) \leq 2\sqrt{p(1-p)}$, and (4.51) is from $\operatorname{sech}(x) \leq \frac{2}{2+x^2}$.

Plugging in the definition for τ_2 into $l(\tau_2)$ we have

$$l(\tau_2) = L_{c,k_{\min}}\left(\frac{2^{\frac{-\tau_2}{d}}}{\sqrt{2\pi e}}\right) = L_{c,k_{\min}}\left(\sqrt{\frac{\varepsilon}{4\pi e c_d}}\right) \tag{4.52}$$

so

$$l(\tau_2) \geq \frac{c^2k_{\min}^2}{128\pi e c_d\frac{1}{\varepsilon} + c^2k_{\min}^2}(1-c) \tag{4.53}$$

which implies

$$\mathbb{E}[T_\varepsilon] \leq \frac{\left(\frac{d}{2}\log_2\frac{c_d}{2\varepsilon} + d + 1\right)\left(128\pi e c_d\frac{1}{\varepsilon} + c^2k_{\min}^2\right)}{(1-c)c^2k_{\min}^2} \tag{4.54}$$

$$\implies \mathbb{E}[T_\varepsilon] = O\left(d\log\frac{1}{\varepsilon} + \left(\frac{1}{\varepsilon k_{\min}^2}\right)d\log\frac{1}{\varepsilon}\right) \tag{4.55}$$

4.6.6 Proof of Proposition 4.3.7

Proof. We first bound $p_1 := \mathbb{P}(Y = 1)$. Recall that for some fixed k, $f(x) = (1 + e^{-kx})^{-1}$. First note that

$$\begin{aligned}\int_a^b f(x)dx &= \int_a^b \frac{1}{k}\frac{ke^{kx}}{1+e^{kx}}dx \\ &= \frac{1}{k}\int_a^b \frac{u'}{u}dx = \frac{1}{k}\int_{u(a)}^{u(b)} \frac{1}{u}dx \\ &= \frac{1}{k}\ln\frac{1+e^{kb}}{1+e^{ka}}.\end{aligned}$$

We have that $\mathbb{P}(Y = 1) = \mathbb{E}[\mathbb{P}(Y = 1|X = x) = \mathbb{E}[f(x)]$. Note that $\forall x$, $(1 + e^{-kx}) \leq 1$. Then,

$$\begin{aligned}p_1 = \mathbb{E}[f(x)] &= \int f(x)p_X(x)dx \\ &= \int_{x\leq 0} f(x)p_X(x)dx + \int_{x>0} f(x)p_X(x)dx \\ &\leq \frac{1}{\sigma_X}\int_{-\infty}^0 f(x)dx + \int_{x>0} f(x)p_X(x)dx \\ &\leq \frac{1}{\sigma_X k}\ln\frac{1+e^{k0}}{1} + \int_{x>0} 1p_X(x)dx \\ &\leq \frac{\ln 2}{\sigma_X k} + \mathbb{P}(X > 0) \leq \frac{\ln 2}{\sigma_X k} + 1 - \frac{1}{e},\end{aligned}$$

where we use $p_X(x) \leq 1/\sigma_X$ and the final inequality follows from $\mathbb{P}(X \leq 0) \geq \frac{1}{e}$ for log-concave X. Using a similar argument it can be shown that $\mathbb{E}[f(x)] \geq 1/e - \ln 2/(\sigma_X k)$. Combining these, we have

$$\frac{1}{e} - \frac{\ln 2}{\sigma_X k} \leq p_1 \leq 1 - \left(\frac{1}{e} - \frac{\ln 2}{\sigma_X k}\right). \tag{4.56}$$

Now we turn to lower bounding $I(X;Y) := H(Y) - H(Y|X)$. The second term can be

written

$$\begin{aligned} H(Y|X) &= \mathbb{E}_X\, H(Y|X=x) \\ &= \int_{-\infty}^{\infty} h_b(f(x)) p_X(x) dx \\ &\leq \frac{1}{\sigma_X} \int_{-\infty}^{\infty} h_b(f(x)) dx. \end{aligned} \tag{4.57}$$

where the inequality follows from Lemma 4.6.1. Since

$$\begin{aligned} &H(Y|X=x) \\ &\quad = -f(x)\log_2 f(x) - (1-f(x))\log_2(1-f(x)) \\ &\quad = \frac{1}{1+e^{-kx}}\log_2(1+e^{-kx}) \\ &\qquad + \frac{e^{-kx}}{1+e^{-kx}}\log_2((1+e^{-kx})/e^{-kx}) \\ &\quad = \frac{1+e^{-kx}}{1+e^{-kx}}\log_2(1+e^{-kx}) \\ &\qquad - \frac{e^{-kx}}{1+e^{-kx}}\log_2(e^{-kx}) \\ &\quad = \log_2(1+e^{-kx}) + \frac{kxe^{-kx}\log_2(e)}{1+e^{-kx}}, \end{aligned}$$

which is an even function, we have (omitting details of the integration)

$$\begin{aligned} H(Y|X) &\leq \frac{2}{\sigma_X}\int_0^{\infty} \log_2(1+e^{-kx}) \\ &\qquad + \frac{kxe^{-kx}\log_2(e)}{1+e^{-kx}} dx \\ &= \frac{2}{\sigma_X}\left(\frac{\pi^2}{12k\log 2} + \frac{\pi^2}{12k\log 2}\right) \\ &= \frac{2}{\sigma_X k} C. \end{aligned} \tag{4.58}$$

For the second term, note that $H(Y=1) = h_b(p)$. The binary entropy function is symmetric about, and monotonically decreasing from $p = 1/2$. Therefore,

$$H(Y) = h_b(p_1) \geq h_b\left(\frac{1}{e} - \frac{\ln 2}{\sigma_X k}\right) \tag{4.59}$$

Combining (4.56) and (4.59) gives the desired result. □

4.6.7 Proof of Lemma 4.6.2

Proof. Since $\log_2 p(\mathbf{W}|\mathbf{y}^{i+1})$ is log-concave, and by Jensen's inequality,

$$\begin{aligned} -h(\mathbf{W}|\mathbf{y}^{i+1}) &= \mathbb{E}_{\mathbf{W}|\mathbf{y}^{i+1}}[\log_2 p(\mathbf{W}|\mathbf{y}^{i+1})] \\ &\leq \log_2 p(\mathbb{E}[\mathbf{W}|\mathbf{y}^{i+1}]|\mathbf{y}^{i+1}) \\ &\leq \log_2 \sup_w p(\mathbf{w}|\mathbf{y}^{i+1}). \end{aligned}$$

Without loss of generality, we may suppose $\mathbb{E}[\mathbf{W}|\mathbf{y}^i] = 0$, and let $\mathbf{V} = \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}^{-\frac{1}{2}}\mathbf{W}$ where and $\mathbf{W} \sim P_{\mathbf{W}|\mathbf{y}^i}$, such that $\mathbb{E}[\mathbf{V}] = 0$ and $\mathbb{E}[\mathbf{V}\mathbf{V}^T] = \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}^{-\frac{1}{2}}\,\mathbb{E}[\mathbf{W}\mathbf{W}^T]\boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}^{-\frac{1}{2}} = \boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}\boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}^{-\frac{1}{2}} = \mathbf{I}$ and therefore $\mathbf{V}$ is isotropic. From [178] we have that $p_V(\mathbf{v}) \leq 2^{8d} d^{\frac{d}{2}}$. From the density of a linear transformation of a random variable we have

$$p_{\mathbf{W}|\mathbf{y}^i}(\mathbf{w}) = \frac{p_V(\boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}^{-\frac{1}{2}} w)}{|\boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}^{\frac{1}{2}}|} \leq \frac{2^{8d} d^{\frac{d}{2}}}{|\boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}|^{\frac{1}{2}}}.$$

Therefore, for our query strategy we have (with $f_{i+1}(\mathbf{W})$ denoting the logistic response model for the query at iteration $i+1$)

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}^{i+1}) &= p(\mathbf{w}|y_{i+1} = y, \mathbf{y}^i) \\ &= \frac{f_{i+1}(\mathbf{W})y + (1 - f_{i+1}(\mathbf{W}))(1-y)}{p(y_{i+1} = y|\mathbf{y}^i)} p(\mathbf{W}|\mathbf{y}^i) \\ &\leq \frac{(1)y + (1-(0))(1-y)}{p(y_{i+1} = y|\mathbf{y}^i)} p(\mathbf{W}|\mathbf{y}^i) \\ &= \frac{1}{p(y_{i+1} = y|\mathbf{y}^i)} p(\mathbf{W}|\mathbf{y}^i) \\ &\leq \frac{1}{p(y_{i+1} = y|\mathbf{y}^i)} \frac{2^{8d} d^{\frac{d}{2}}}{|\boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}|^{\frac{1}{2}}} \\ \implies \sup_w p(\mathbf{w}|\mathbf{y}^{i+1}) &\leq \frac{1}{p(y_{i+1} = y|\mathbf{y}^i)} \frac{2^{8d} d^{\frac{d}{2}}}{|\boldsymbol{\Sigma}_{\mathbf{W}|\mathbf{y}^i}|^{\frac{1}{2}}}, \end{aligned}$$

which implies

$$\log_2 \sup_w p(\mathbf{w}|\mathbf{y}^{i+1}) \leq 8d + \frac{d}{2}\log_2 d - \frac{1}{2}\log_2|\mathbf{\Sigma}_{\mathbf{W}|\mathbf{y}^i}| - \log_2(p(y_{i+1} = y|\mathbf{y}^i)),$$

and hence

$$\begin{aligned}
h(\mathbf{W}|\mathbf{y}^{i+1}) &\geq \frac{1}{2}\log_2|\mathbf{\Sigma}_{\mathbf{W}|\mathbf{y}^i}| + \log_2(p(y_{i+1} = y|\mathbf{y}^i)) \\
&\quad - \left(8d + \frac{d}{2}\log_2 d\right) \\
&\geq \frac{1}{2}\log_2((2\pi e)^d|\mathbf{\Sigma}_{\mathbf{W}|\mathbf{y}^i}|) \\
&\quad - \frac{1}{2}\log_2(2\pi e)^d + \log_2(p(y_{i+1} = y|\mathbf{y}^i)) \\
&\quad - \left(8d + \frac{d}{2}\log_2 d\right) \\
&\geq h(\mathbf{W}|\mathbf{y}^i) + \log_2(p(y_{i+1} = y|\mathbf{y}^i)) \\
&\quad - \left(8d + \frac{d}{2}\log_2(2\pi e d)\right) \quad \text{from (4.12).}
\end{aligned}$$

For equiprobable queries ($p(y_{i+1} = y|\mathbf{y}^i) = 1/2$), and so we have

$$h(\mathbf{W}|\mathbf{y}^i) - h(\mathbf{W}|\mathbf{y}^{i+1}) \leq \gamma(d). \tag{4.60}$$

where $\gamma(d) = 8d + \frac{d}{2}\log_2(2\pi e d) + 1$.

To obtain the other direction, let $h_y^i = h(\mathbf{W}|Y_{i+1} = y, \mathbf{y}^i)$, $y_m = \arg\min_{y\in\{0,1\}} h_y^i$, $y_M = 1 - y_m$. Note that $h_{y_M}^i \geq h_{y_m}^i$. We have

$$\begin{aligned}
h(\mathbf{W}|Y_{i+1}, \mathbf{y}^i) &= \frac{1}{2}h_m^i + \frac{1}{2}h_M^i \\
&\geq \frac{1}{2}(h(\mathbf{W}|\mathbf{y}^i) - \gamma(d)) + \frac{1}{2}h_M^i \\
&\geq \frac{1}{2}(h(\mathbf{W}|\mathbf{y}^i) - \gamma(d)) + \frac{1}{2}h(\mathbf{W}|\mathbf{y}^{i+1})
\end{aligned}$$

where the first inequality follows from (4.60) and the second inequality follows from the definition of h_M. From the non-negativity of mutual information, we have that $h(\mathbf{W}|Y_{i+1}, \mathbf{y}^i) \leq$

$h(\mathbf{W}|\mathbf{y}^i)$, implying

$$h(\mathbf{W}|\mathbf{y}^i) \geq \frac{1}{2}(h(\mathbf{W}|\mathbf{y}^i) - \gamma(d)) + \frac{1}{2}h(\mathbf{W}|\mathbf{y}^{i+1})$$

$$h(\mathbf{W}|\mathbf{y}^i) - h(\mathbf{W}|\mathbf{y}^{i+1}) \geq -\gamma(d) \tag{4.61}$$

Combining (4.61) with (4.60) we have the desired result. □

CHAPTER 5

MEASUREMENT SELECTION AND APPLICATIONS

In this chapter we revisit the problem of intelligently selecting a set of measurements to improve our ability to perform estimation. Suppose we had some prior knowledge or existing estimate of the signal to be measured, and we would like to adapt or "tune" future measurements to leverage this information in order to improve our estimation ability. This goal forms an important sub-problem of general adaptive sensing and can be considered of as an *exploit* phase which follows an *explore* phase.

We start with the familiar linear sensing model and discuss the method of [179], which ties up some questions from Chapter 2, where we discussed constrained adaptive sensing. We then show that this method can be used more generally by giving some examples, including application to generalized linear models and pairwise comparisons.

5.1 Motivation

To motivate this discussion, first we consider again the linear constrained sensing setting. Specifically, suppose that we may take $m < n$ noisy linear measurements of a n-dimensional k-sparse signal;

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z},$$

where $\|\mathbf{x}\|_0 = k$ and $z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Further suppose that each row of $\mathbf{A}$, denoted by $\mathbf{a}_i^T$, is drawn from a fixed collection of size M, i.e., it is highly constrained;

$$\mathbf{a}_i^T \in \{\mathbf{u}_1^T; \mathbf{u}_2^T; \cdots; \mathbf{u}_M^T\} =: \mathbf{U} \in \mathbb{R}^{M \times n}.$$

Suppose we are given the sparse support of $\mathbf{x}$, $\Lambda := \text{supp}(x)$, we set $\hat{x}_\Lambda = A_\Lambda^+ y$, $\hat{x}_{\Lambda^c} = 0$. In this case we have

$$\mathbb{E}\|\mathbf{x}|_\Lambda - \hat{\mathbf{x}}(\mathbf{y})|_\Lambda\|_2^2 = \sigma^2 \,\text{Tr}\left((\mathbf{A}_\Lambda^T \mathbf{A}_\Lambda)^{-1}\right) = \sum_{i=1}^{k} \frac{\sigma^2}{\lambda_i(\mathbf{A}_\Lambda^T \mathbf{A}_\Lambda)} \leq \frac{k\sigma^2}{\lambda_{\min}(\mathbf{A}_\Lambda^T \mathbf{A}_\Lambda)}.$$

The trace may be recognized as resembling the *A-optimal experimental design* objective. The right-hand side of the inequality shows that the mean squared error is dominated by the smallest eigenvalue of the Gram matrix.

Measurement selection problem.— Given $\Lambda = \text{supp}(\mathbf{x})$ and hence $\mathbf{U}_\Lambda$, choose the m rows of $\mathbf{A}$ from the rows of $\mathbf{U}$ so as to minimize

$$\text{Tr}\left((\mathbf{A}_\Lambda^T \mathbf{A}_\Lambda)^{-1}\right) = \text{Tr}\left(\sum_{\ell=1}^{n} w_\ell \, \mathbf{u}_{\ell\Lambda}\mathbf{u}_{\ell\Lambda}^T\right)^{-1} \tag{5.1}$$

where $w_\ell \in \{0, 1, \cdots, b\}$ is a weight assigned to each measurement and $\sum_\ell w_\ell \leq m$. We let $b \geq 1$ refer to an upper bound on the number of times each measurement may be duplicated, depending on whether we wish to allow measurement repetition or not. Whether this is possible or desirable depends on the specific problem at hand.

Since this problem involves optimizing over a discrete domain, it is not expected to be tractable. Instead, we relax the problem to a *continuous* version, introducing weights $\widehat{w}_\ell \in [0, b]$ for $\ell \in [n]$, and minimize, subject to $\sum_\ell \widehat{w}_\ell = m$,

$$\text{Tr}\left(\sum_{\ell=1}^{n} \widehat{w}_\ell \, \mathbf{u}_{\ell\Lambda}\mathbf{u}_{\ell\Lambda}^T\right)^{-1} = \text{Tr}\left(\mathbf{U}_\Lambda^T \, \text{diag}(\widehat{\mathbf{w}})\mathbf{U}_\Lambda\right)^{-1}. \tag{5.2}$$

This optimization problem is quite advantageous because it is a tractable convex program, and often works empirically [34]. It is also somewhat well-known in the experimental design literature. Unfortunately, it is difficult to relate (5.2) to our original problem. First, it is not guaranteed that the continuous relaxation is close to the discrete optimization problem. That is, we must somehow *round* the "continuously weighted" $\widehat{w}_\ell$ to m discrete measurements. Second, even in the relaxed setting of (5.2), it is unclear how small or large the objective be for an optimal solution and what properties of the collection $\mathbf{U}$ affect those possible solutions. We aim to investigate these two concerns here.

5.2 Rounding

Recently, in [179], Allen-Zhu et al. introduce a novel algorithm to approximately choose a subset of measurements which to minimize the A-optimality criterion (among other common criteria). This is done by first solving the continuous, convex, relaxation (5.2), then giving a procedure and guarantee on rounding the solution in order to estimate the discrete problem with integrality constraints (5.1). Importantly, they show their algorithm gives an $O(1+\epsilon)$-approximation to the desired objective provided the measurement budget is at least $O(k/\epsilon^2)$. We give an outline of their methods and results here, then give some applications relevant to the material in this thesis.

5.2.1 Preliminaries

We assert that the objective of (5.1) satisfies a couple of important properties which will enable the subsequent analysis. Let $\mathbf{G} = \mathbf{A}^T\mathbf{A}$ be a Gram matrix. First, note that the *normalized A-optimal objective*

$$f_A(\mathbf{G}) := \frac{1}{k}\operatorname{Tr}(\mathbf{G}^{-1})$$

is *monotonic* according to the Loewner order, where for any $\mathbf{G}_1, \mathbf{G}_2 \in S_k^+$ (the k dimensional positive semi-definite cone),

$$\mathbf{G}_1 \preceq \mathbf{G}_2 \implies f_A(\mathbf{G}_1) \geq f_A(\mathbf{G}_2).$$

It is also *reciprocal sub-linear*, since for any $\mathbf{G} \in S_k^+$ and any $t \in (0,1)$, we have

$$f_A(t\mathbf{G}) \leq \frac{1}{t} f_A(\mathbf{G}).$$

In fact it is linear and this holds with equality, although we will not need that. It will also be convenient to define a few sets; denoting the spaces of possible continuous and discrete

weight vectors;

$$S_m := \{\mathbf{s} \in \{0,1\}^M \,:\, \sum_\ell s_\ell \le m\},$$

$$S_m^b := \{\mathbf{s} \in \{0,\dots,b\}^M \,:\, \sum_\ell s_\ell \le m\},$$

$$C_m^b := \{\mathbf{s} \in [0,b]^M \,:\, \sum_\ell s_\ell \le m\}.$$

Let $F : C_m^b \to \mathbb{R}$ be the objective for a particular weight vector, i.e.,

$$F(s) := f_A\left(\sum_{\ell=1}^M s_\ell \mathbf{u}_\ell \mathbf{u}_\ell^T\right).$$

Let $\boldsymbol{\pi}$ refer to the solution to the continuous problem (5.2), which we assume can be efficiently solved to arbitrary accuracy. The goal will be the following;

$$\text{Find an } \hat{\mathbf{s}} \in S_m \text{ such that } \|\hat{\mathbf{s}}\|_1 = m,\ \sum_\ell \hat{s}_\ell \mathbf{u}_\ell^T \mathbf{u}_\ell \succeq (1-3\epsilon) \sum_\ell \pi_\ell \mathbf{u}_\ell^T \mathbf{u}_\ell \tag{5.3}$$

for some constant ϵ. Note that we can without loss of generality assume that

$$\sum_\ell \pi_\ell \mathbf{u}_\ell \mathbf{u}_\ell^T = \mathbf{I}.$$

To see this, note that any collection $\{\mathbf{u}'_\ell\}_\ell$ where

$$\sum_\ell \pi_\ell \mathbf{u}'_\ell \mathbf{u}'^T_\ell = \mathbf{U}^T \mathrm{diag}(\boldsymbol{\pi}) \mathbf{U} =: \mathbf{V}$$

for positive-definite $\mathbf{V}$ can be brought into isotropic position by setting $\mathbf{u}_\ell = \mathbf{V}^{-1/2}\mathbf{u}'_\ell$. Then,

$$\sum_\ell \pi_\ell u_\ell u_\ell^T = \sum_\ell \pi_\ell \mathbf{V}^{-1/2}\mathbf{u}'_\ell \mathbf{u}'^T_\ell \mathbf{V}^{-T/2} = \mathbf{V}^{-1/2}(\sum_\ell \pi_\ell \mathbf{u}'_\ell \mathbf{u}'^T_\ell)\mathbf{V}^{-T/2} = \mathbf{V}^{-1/2}\mathbf{V}\mathbf{V}^{-T/2} = \mathbf{I}.$$

Now, if we could solve the isotropic problem of finding an s such that

$$\sum_\ell \hat{s}_\ell \mathbf{u}'_\ell \mathbf{u}'^T_\ell \succ c \sum_\ell \pi_\ell \mathbf{u}'_\ell \mathbf{u}'^T_\ell = c\mathbf{I},$$

we would likewise have

$$\sum_{\ell} \hat{s}_\ell \mathbf{u}_\ell \mathbf{u}_\ell^T = \mathbf{V}^{-T/2}(\sum \hat{s}_\ell \mathbf{u}'_\ell \mathbf{u}'^T_\ell)\mathbf{V}^{-1/2} \succ \mathbf{V}^{-T/2}(c\mathbf{I})\mathbf{V}^{-1/2} = c\mathbf{V}.$$

To see that a solution $\hat{\mathbf{s}}$ to problem (5.3) would imply a result on $F(s)$, note that by monotonicity and reciprocal sub-linearity,

$$F(\hat{\mathbf{s}}) = f_A(\sum \hat{s}_\ell \mathbf{u}_\ell \mathbf{u}_\ell^T) \leq f_A((1-3\epsilon)\mathbf{I}) \leq \frac{1}{1-3\epsilon} F(\boldsymbol{\pi}) \leq (1+6\epsilon)F(\boldsymbol{\pi})$$

for sufficiently small ϵ (say $\epsilon \leq 1/3$).

The remaining analysis involves giving an efficient procedure to solve the problem (5.3).

5.2.2 Primary results of [179]

Theorem 5.2.1 (Rounding, Thm 2.1 in [179])**.** *Suppose $\epsilon \in (0, 1/3]$, $n \geq m \geq 5k/\epsilon^2$, $b \in \{1, 2, \ldots, m\}$, and $f : S_m^+ \to \mathbb{R}$ is monotone and reciprocal sub-linear. Let $\boldsymbol{\pi} \in C_m^b$ be any "fractional" solution to* (5.2) *such that $F(\boldsymbol{\pi}) < \infty$. Then, there is a polynomial time algorithm which rounds $\boldsymbol{\pi}$ to an integral solution $\mathbf{s}_b$ satisfying*

$$F(\mathbf{s}_b) \leq (1+6\epsilon)F(\boldsymbol{\pi}).$$

Theorem 5.2.2 (Selection, Thm 1.4 in [179])**.** *Suppose $\epsilon \in (0, 1/3]$, $n \geq m \geq 5k/\epsilon^2$, $b \in \{1, \ldots, k\}$, and assume $f : S_m^+ \to \mathbb{R}$ is monotone, reciprocal sub-linear, and is such that problem $\min_{\mathbf{s} \in S_k^b} F(\mathbf{s}) < \infty$ can be solved in polynomial time. Then, there is a polynomial-time algorithm that outputs $s_b \in S_k^b$ satisfying*

$$F(\mathbf{s}_b) \leq (1+6\epsilon) \min_{\mathbf{s} \in S_k^b} F(\mathbf{s}).$$

In other words, as long as we allow enough measurements, an $O(1+\epsilon)$ approximation to the optimal continuous relaxation is easy to obtain. This is *also* an approximation the best possible discrete m measurement selection because by definition the continuous problem (5.2) has a solution which is no worse than that of the optimal discrete selection in (5.1). Note that in [179, Theorem 1.4] the additive constant is given as 8ϵ whereas we have 6ϵ be-

cause we do not consider the error when solving convex program (5.2). Instead, we take for granted that efficient, highly accurate solvers exist.

5.2.3 Proof sketch

The results of [179] are shown through the framework of *regret minimization*. This is a commonly used technique in many learning problems and is sometimes called *optimism in the face of uncertainty* [180]. Additionally, the algorithm will use *swapping* to optimize the set of measurements, where at each iteration we drop a measurement in favor of some other. Swapping algorithms are widely used experimental design problems, e.g., the classical Fedorov exchange algorithm [181, 182].

We must first reduce the problem to a set of simple steps which is in some way close to the original hard problem. At a high level, the strategy can be thought of as a game between two players, A and B, where the first chooses "good" sets of measurements and the second chooses "bad" signals. The idea is to produce a set of measurements which is acceptable even on the worst possible signal (i.e., $\lambda_{\min}(\mathbf{A}^T\mathbf{A}) \geq c$).

1. Player A is an adversary who wishes to worsen player B's score by choosing, at each round t, an unfavorable state $\mathbf{F}_t$, which is a $p \times p$ matrix. This state is hidden from the other player until they play an action. We further assume that player A is limited to rank-2 moves of the form $\mathbf{F}_t = \mathbf{w}_t\mathbf{w}_t^T - \mathbf{v}_t\mathbf{v}_t^T$ for some vectors $\mathbf{w}$ and $\mathbf{v}$. If we set $\mathbf{F}_t = \sum_{\ell \in S_t} \mathbf{u}_\ell \mathbf{u}_\ell^T$ given a particular subset $S_t \subset [n]$, these adversary moves can be thought of as *swaps*, or reducing the weight on one measurement while increasing the weight of another. S_0 is initialized with an arbitrary set of m vectors. This setup gives sufficient freedom to optimize the selection of measurements while also minimizing algorithmic complexity.
2. At each round t, player B will choose an action $\mathbf{B}_t$, which is also a $p \times p$ matrix with the additional constraints $\mathbf{B}_t \succeq 0$ and $\mathrm{Tr}(\mathbf{B}_t) \leq 1$, only after which state $\mathbf{F}_t$ is revealed.

Player B will try the minimize the particular score function

$$\sum_{t'=0}^{t-1} \langle \mathbf{F}_{t'}, \mathbf{B}_{t'} \rangle$$

where $\langle \mathbf{F}_t, \mathbf{B}_t \rangle = \mathrm{Tr}(\mathbf{F}_t^T \mathbf{B}_t)$. This player's final score will be their regret relative to the best possible action $\mathbf{B}$ they could have chosen if they had advance knowledge of the $\mathbf{F}_t$ sequence;

$$R_T(\mathbf{B}_{0..T-1}) = \sum_{t=0}^{T-1} \langle \mathbf{F}_t, \mathbf{B}_t \rangle - \min_{\mathbf{B}} \sum_t \langle \mathbf{F}_t, \mathbf{B} \rangle.$$

Note that $\min_{\mathbf{B}} \sum_t \langle \mathbf{F}_t, \mathbf{B} \rangle = \lambda_{\min}(\sum_t \mathbf{F}_t)$. If we could provide an upper bound on the regret R_T, we would immediately have a lower bound on the minimum eigenvalue as well as a series of actions $\mathbf{F}_0 \to \ldots \to \mathbf{F}_{T-1}$ to produce $\mathbf{F}_{T-1} = \sum_{\ell \in S_{T-1}} \mathbf{u}_\ell \mathbf{u}_\ell^T$. Furthermore, since the total number non-zero entries in S_t does not change at each step, the algorithm has chosen exactly m measurements. This procedure is conceptually and algorithmically simple and if the number of iterations, T, is polynomial, it will be efficient as well. The proof of the approximation guarantee will involve bounding regret with T at most k/ϵ, i.e., a small number of steps. Compare this to the original combinatorial problem, which is often NP-hard (depending on the choice of experimental design function).

The authors of [179] propose solving the regret minimization problem with an $\ell_{1/2}$ follow-the-regularized-leader (FTRL) strategy. Specifically, at each step t, player B's strategy is given by

$$\mathbf{B}_t = \arg\min_{\mathbf{B}'} \psi(\mathbf{B}') - \psi(\mathbf{B}_{t-1}) - \langle \nabla\psi(\mathbf{B}_{t-1}), \mathbf{B}' - \mathbf{B}_{t-1} \rangle + \gamma \langle \mathbf{F}_{t-1}, \mathbf{B}' \rangle$$

where $\psi(\mathbf{B}) := -2\,\mathrm{Tr}(\mathbf{B}^{1/2})$ is a type of regularizer and $\gamma = \sqrt{k}/\epsilon$ controls the rate that player B learns. It turns out that this problem has the closed-form solution

$$\mathbf{B}_t = \left(c_t \mathbf{I} + \gamma \sum_{\ell \in S_0} \mathbf{u}_\ell \mathbf{u}_\ell^T + \gamma \sum_{t'=0}^{t-1} \mathbf{F}_t \right)^{-2}, \tag{5.4}$$

where c_t is chosen such that

$$c_t \mathbf{I} + \gamma \sum_{\ell \in S_0} \mathbf{u}_\ell \mathbf{u}_\ell^T + \gamma \sum_{t'=0}^{t-1} \mathbf{F}_t \succ 0 \quad \text{and} \quad \mathrm{Tr}(\mathbf{B}_t) = 1.$$

This constant (which is unique) can be found efficiently by e.g., binary search. The next result can provide a bound on the regret.

Lemma 5.2.3 (Lem 2.5+2.7 in [179]). *Suppose* $\mathbf{F}_t = \mathbf{w}_t \mathbf{w}_t^T - \mathbf{v}_t \mathbf{v}_t^T$, *that* $\mathbf{B}_t$ *is given by* (5.4), $\mathbf{Z}_0 := \sum_{\ell \in S_0} \mathbf{w}_\ell \mathbf{w}_\ell^T$, *and* $\gamma \langle \mathbf{B}_t^{1/2}, \mathbf{v}_t \mathbf{v}_t^T \rangle < 1/2$. *Then,*

$$-\sum_{t=0}^{T-1} \langle \mathbf{F}_t, \mathbf{U} \rangle \leq \sum_{t=0}^{T-1} \left(-\frac{\langle \mathbf{B}_t, \mathbf{w}_t \mathbf{w}_t^T \rangle}{1 + 2\gamma \langle \mathbf{B}_t^{1/2}, \mathbf{w}_t \mathbf{w}_t^T \rangle} + \frac{\langle \mathbf{B}_t, \mathbf{v}_t \mathbf{v}_t^T \rangle}{1 - 2\gamma \langle \mathbf{B}_t^{1/2}, \mathbf{v}_t \mathbf{v}_t^T \rangle} \right) + \frac{2\sqrt{p} + \gamma \langle \mathbf{Z}_0, \mathbf{U} \rangle}{\gamma}.$$

The proof of this Lemma is very technical and we refer the reader to [179, Appendix]. Recall that the vectors $\mathbf{w}_t$ and $\mathbf{v}_t$ chosen in Lemma 5.2.3 are Player A's choice of the two $\mathbf{u}_i$ and $\mathbf{u}_j$ to swap. It remains to show that at any stage, there in fact exists two such vectors satisfying the assumptions of Lemma 5.2.3, and to provide upper bounds for each of the result's terms. The following estimates can be given; for any $S \subset [n]$ where $|S| = m$, and assuming $\lambda_{\min}(\sum_{\ell \in S_T} \mathbf{u}_\ell \mathbf{u}_\ell^T) \leq 1 - 3\epsilon$,

$$\min_{i \in S : 2\gamma \langle \mathbf{B}^{1/2}, \mathbf{u}_i \mathbf{u}_i^T \rangle < 1} \frac{\langle \mathbf{B}, \mathbf{u}_i \mathbf{u}_i^T \rangle}{1 - 2\gamma \langle B^{1/2}, \mathbf{u}_i \mathbf{u}_i^T \rangle} \leq \frac{1 - \epsilon}{m}$$

and

$$\max_{j \in [n] \setminus S} \frac{\langle \mathbf{B}, \mathbf{u}_j \mathbf{u}_j^T \rangle}{1 - 2\gamma \langle \mathbf{B}^{1/2}, \mathbf{u}_j \mathbf{u}_j^T \rangle} \geq \min_{i \in S : 2\gamma \langle \mathbf{B}^{1/2}, \mathbf{u}_i \mathbf{u}_i^T \rangle < 1} \frac{\langle \mathbf{B}, \mathbf{u}_i \mathbf{u}_i^T \rangle}{1 - 2\gamma \langle \mathbf{B}^{1/2}, \mathbf{u}_i \mathbf{u}_i^T \rangle} + \frac{\epsilon}{m}.$$

Finally, to show there exists an i such that $2\gamma \langle \mathbf{B}_t^{1/2}, \mathbf{u}_i \mathbf{u}_i^T \rangle < 1$ so the denominator is not infinity, consider that if this were not the case, since if $|S| = m$,

$$\sum_{\ell \in S} 2\gamma \langle \mathbf{B}_t^{1/2}, \mathbf{u}_\ell \mathbf{u}_\ell^T \rangle \geq m.$$

However, using $m \geq 5k/\epsilon^2$, it can be shown that

$$2\gamma \langle \mathbf{B}_t^{1/2}, \sum_{\ell \in S} \mathbf{u}_\ell \mathbf{u}_\ell^T \rangle \leq 2k + 2\gamma\sqrt{k} = 2k + 2(\sqrt{k}/\epsilon)\sqrt{k} \leq 2\epsilon^2/5 + 2\epsilon m/5 < m.$$

These bounds, combined with Lemma 5.2.3, and $T \geq m/\epsilon$, eventually give

$$-\lambda_{\min}\Big(\sum_{\ell \in S_T} \mathbf{u}_\ell \mathbf{u}_\ell^T\Big) \leq \sum_{t=0}^{T-1} -\frac{\epsilon}{m} + \frac{2\sqrt{k}}{\gamma} = -\frac{T\epsilon}{m} + 2\epsilon \leq -1 + 2\epsilon,$$

which implies $\lambda_{\min}\big(\sum_{j \in S_T} \mathbf{u}_j \mathbf{u}_j^T\big) \geq 1 - 3\epsilon$ as desired.

5.3 Application to generalized linear models

We consider generalized linear models (GLMs) due to their application to quantized sensing and because they are well-studied and particularly convenient to work with. In this section we introduce the following notation: Let Ξ be the set of possible experiments and let each $\xi_i \in \Xi$ represent an abstract experimental setup. The ith row of the measurement matrix $\mathbf{A}$ will be related through the function

$$\mathbf{a}_i^T = h(\xi_i).$$

This indirection in notation is more standard in the GLM literature and allows the possibility that the sensing vectors depend on some arbitrary experimental parameters in a non-linear way. If the set of measurements and experimental parameters are both discrete there is otherwise no difference to the analysis. Given a set of parameters $\mathbf{x}$ and an experimental setup ξ, a GLM consists of (see e.g., [183]) a conditional probability distribution (usually of an exponential family) for the response $y_i \sim p(y|\mathbf{x}, \xi)$ which has three properties (i) a response y_i such that $\mathbb{E}\, y_i = \mu_i$, (ii) a link function $g(\mu_i) = \eta_i$, and (iii) a linear predictor $\eta_i = h(\xi)^T \mathbf{x}$. Together, $g(\mathbb{E}\, y_i) = h(\xi)^T \mathbf{x}$.

Remark. *A very common example of a GLM is the* logistic regression model *in which*

$$y_i \in \{0, 1\}, \quad \eta_i = \langle \mathbf{a}_i, \mathbf{x} \rangle, \quad \mu_i = g^{-1}(\eta_i) = \frac{1}{1 + \exp(-\langle \mathbf{a}_i, \mathbf{x} \rangle)}.$$

Here, μ can also be thought of as $\mu_i = \mathbb{P}(y_i = 1) = g^{-1}(\langle \mathbf{a}_i, \mathbf{x} \rangle)$. We will study the specifics of the logistic noise case shortly.

The goal here by imposing a GLM is to allow estimation of the parameters $\mathbf{x}$ from a

set of responses $\{y_i\}_i$. Frequently, maximum likelihood estimation (MLE) is used. The exact form of the MLE depends on the likelihood function $l(\mathbf{x}|y) = p(y|\mathbf{x})$, but subject to regularity conditions [184], we can describe the asymptotic behavior with one other piece of information, $\mathrm{Var}(y_i)$. Specifically, as $m \to \infty$ the solution to the MLE $\hat{\mathbf{x}}_{\mathrm{MLE}}$ satisfies:

$$\hat{\mathbf{x}}_{\mathrm{MLE}} \rightsquigarrow \mathcal{N}(\mathbf{x}, \mathbf{M}(\mathbf{x})^{-1}).$$

$M(\mathbf{x})$ is the *information matrix* which can be written as

$$\mathbf{M}(\mathbf{x}) = \sum_i \mathbf{M}(\xi_i; \mathbf{x}) = \sum_i \frac{1}{w(\xi_i)} h(\xi_i) h(\xi_i)^T = h(\xi)^T \mathbf{W}^{-1}(\xi) h(\xi),$$

where $\mathbf{W}(\xi) = \mathrm{diag}(w(\xi_1), \ldots, w(\xi_m))$, and $w(\xi_i)$ are a set of weights given by [184]

$$w(\xi_i) = \mathrm{Var}(y_i)(g'(\mu_i))^2 = \mathrm{Var}(y_i)(\partial \mu_i / \partial \eta_i)^{-2}.$$

The inverse of the information matrix also gives a lower bound on the covariance of any unbiased estimator of $\mathbf{x}$ via the *Cramér–Rao lower bound*;

$$\mathrm{Cov}(\hat{\mathbf{x}}) \succeq \mathbf{M}^{-1}(\xi) = (h(\xi)^T \mathbf{W}(\xi)^{-1} h(\xi))^{-1}.$$

Remark. *In the familiar linear Gaussian-noise setting where* $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, $\mu_i = \mathbf{a}_i^T \mathbf{x}$, *and* $\mathrm{d}\mu_i / \mathrm{d}\eta_i = 1$, *this becomes the following, with equality corresponding to recovery using the pseudo-inverse;*

$$\mathrm{Cov}(\hat{\mathbf{x}})) = (h(\xi)^T h(\xi))^{-1} \implies \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|^2 = \sigma^2 \,\mathrm{Tr}[(\mathbf{A}^T \mathbf{A})^{-1}].$$

As mentioned previously, in optimal experimental design one seeks to minimize a summary functional of the inverse of the information matrix, such as the trace [181]. While the information matrix does not in general (in non-linear or non-Gaussian cases) provide an upper bound on estimation error, this can provide a useful heuristic.

Note that $\mathbf{W}(\xi)$ and $\mathbf{M}(\xi)$ depend implicitly on the true parameters $\mathbf{x}$ and hence cannot be directly estimated. However, if there exist upper and lower bounds valid for any parameter

of interest, e.g., if there is a set C such that for any $\mathbf{x} \in C$ we have

$$L_C \leq w(\xi_i; \mathbf{x}) \leq U_C \tag{5.5}$$

then

$$L_C(h(\xi)^T h(\xi))^{-1} \preceq \mathbf{M}^{-1}(\xi) \preceq U_C(h(\xi)^T h(\xi))^{-1}.$$

As m grows very large, we expect

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|^2 \approx kU_C \operatorname{Tr}[(\mathbf{A}^T\mathbf{A})^{-1}].$$

Thus, at least heuristically, we can use the results of [179] to choose the set of $\{\mathbf{a}_i\}_i$ from the available measurements to improve our estimate.

Corollary 5.3.1 (of Theorem 5.2.2)**.** *If $m \geq 5k/\epsilon^2$ there a polynomial-time algorithm which selects a set $S \subset [M]$ with $|S| = m$ such that the maximum likelihood estimate $\hat{\mathbf{x}}_{MLE}$ in* (5.8) *asymptotically sastisfies*

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{x}}_{MLE} - \mathbf{x}\|^2 &= O\Big(kU_C \operatorname{Tr}\Big[\Big(\sum_{i\in S} \mathbf{u}_i\mathbf{u}_i^T\Big)^{-1}\Big]\Big) \\ &= O\Big(kU_C(1+6\epsilon) \inf_{|S|=m} \operatorname{Tr}\Big[\Big(\sum_{i\in S} \mathbf{u}_i\mathbf{u}_i^T\Big)^{-1}\Big]\Big).\end{aligned}$$

5.3.1 Logistic regression

As a concrete example of a GLM, we consider the logistic regression model with fidelity constant κ, where

$$\mu_i = \frac{1}{1 + \exp(-\kappa\eta_i)}$$

and thus

$$w(\xi_i; \mathbf{x}) = \frac{2}{k^2}(1 + \cosh(-\kappa\langle \mathbf{a}_i, \mathbf{x}\rangle)), \tag{5.6}$$

which is an increasing function in $\langle \mathbf{a}_i, \mathbf{x}\rangle$. Let C be the set of $\mathbf{x}$ such that $\|\mathbf{x}\| \leq R$ for some $R > 0$. Then we have

$$L_C = \frac{2}{\kappa^2} \leq w(\xi_i, \mathbf{x}) \leq \frac{2}{\kappa^2}(1 + \cosh(-\kappa R)) = U_C.$$

Suppose $\|a_i\| = 1$ for all i. From (5.6), we obtain

$$\frac{4}{\kappa^2}\Big(\sum_i \mathbf{a}_i\mathbf{a}_i^T\Big)^{-1} \preceq \mathbf{M}^{-1}(\xi) \preceq \frac{2}{\kappa^2}(1+\cosh(-\kappa R))\Big(\sum_i \mathbf{a}_i\mathbf{a}_i^T\Big)^{-1}.$$

Note that the function $(1+\cosh(-\gamma))/\gamma^2$ is minimized at $\gamma \approx 2.4$, so this upper bound can be tightened in situations where we have control over parameters R or κ. Setting $\kappa = 2.4/R$, this yields the following estimate of the error, relative to the problem scale:

$$\frac{\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|^2}{R^2} = O(\mathrm{Tr}[(\mathbf{A}^T\mathbf{A})^{-1}]).$$

Alternatively, setting $R = 2.4/\kappa$, this can be written in terms of κ as

$$\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|^2 = O(\kappa^{-2}\,\mathrm{Tr}[(\mathbf{A}^T\mathbf{A})^{-1}]).$$

As in Corollary (5.3.1), if $m \geq 5k/\epsilon^2$ then there an efficient algorithm which selects an $S \subset [M]$ with $|S| = m$ such that

$$\begin{aligned}\frac{\mathbb{E}\|\hat{\mathbf{x}}_{\mathrm{MLE}} - \mathbf{x}\|^2}{R^2} &= O\Big(k\,\mathrm{Tr}\Big[\Big(\sum_{i\in S}\mathbf{u}_i\mathbf{u}_i^T\Big)^{-1}\Big]\Big) \\ &= O\Big(k(1+6\epsilon)\inf_{|S|=m}\mathrm{Tr}\Big[\Big(\sum_{i\in S}\mathbf{u}_i\mathbf{u}_i^T\Big)^{-1}\Big]\Big).\end{aligned}$$

A more rigorous upper bound on error will be given in the context of a related but slightly different model.

5.4 Application to estimation with pairwise comparisons

We now consider application to a logistic model which is common in psychometric literature. See Chapter 3 for more background and details (note that the notation in this section is slightly different). This model is very similar to the GLM case, except with introduction of the threshold, which will make analysis somewhat more challenging.

We investigate the problem of estimating a point $\mathbf{x}^* \in \mathbb{R}^k$ given noisy quantized observations $y_1, \ldots, y_m \subset \{-1, 1\}^m$ where each $y_i = 1$ means "$\mathbf{x}^*$ is closer to $\mathbf{p}_i$ than it is to $\mathbf{q}_i$" for some $\mathbf{p}_i, \mathbf{q}_i \in Q \subset \mathbb{R}^k$. Each measurement i is equivalently identified with a *direction*

and *threshold* $(\mathbf{a}_i, \tau_i)$ given by

$$\mathbf{a}_i = 2(\mathbf{p}_i - \mathbf{q}_i), \quad \tau_i = \|\mathbf{p}_i\|^2 - \|\mathbf{q}_i\|^2.$$

Note that we do not in general assume $\|\mathbf{a}_i\| = 1$. In the absence of noise, we could write

$$y_i = \text{sign}(\|\mathbf{x}^* - \mathbf{q}\|^2 - \|\mathbf{x}^* - \mathbf{p}\|^2) = \text{sign}(\mathbf{a}_i^T \mathbf{x}^* - \tau_i).$$

Thus this model is naturally understood as a sequence of queries each yielding information about which side of a hyperplane the target signal lies on. Hence, the ability to recover a signal will depend greatly on the geometry. In the case of noise, we will relax this notion, but it will still be helpful to think about the hyperplanes instead of the pairs which generated them. The total collection of possible measurements is denoted by $\overline{\Omega}$ where $|\overline{\Omega}| = M$. Our goal will be to understand the error in estimating $\mathbf{x}^*$ for a given Ω and to select $\Omega \subset \overline{\Omega}$ with $m = |\Omega| < M$ to improve our estimation error.

5.4.1 Maximum likelihood estimation

Suppose that we have full control over the measurement selection and are able to choose some $\Omega \subset \overline{\Omega}$. Let $\mathbf{A}_\Omega$ be the matrix formed from the measurement directions in a particular Ω concatenated as rows and let τ_Ω be the corresponding sub-vector of thresholds. We observe the following;

$$\mathbf{y} = \text{sign}(\boldsymbol{\kappa}_\Omega \circ (\mathbf{A}_\Omega \mathbf{x}^* - \boldsymbol{\tau}_\Omega) + \upsilon), \qquad \mathbf{x}^* \in \mathbb{R}^k. \tag{5.7}$$

where υ is independent noise and $\boldsymbol{\kappa}_\Omega$ is a vector of possibly measurement-dependent fidelity parameters. We will assume a Bernoulli noise model, i.e., for each $i \in \Omega$,

$$y_i = \begin{cases} +1 & \text{w.p.} \quad f(\kappa_i(\mathbf{a}_i^T \mathbf{x}^* - \tau_i)) \\ -1 & \text{w.p.} \quad 1 - f(\kappa_i(\mathbf{a}_i^T \mathbf{x}^* - \tau_i)), \end{cases}$$

where $f : \mathbb{R} \to [0, 1]$ is a link function (for example, the logistic function).

The log-likelihood for this model is given by

$$F_{\Omega,y}(\mathbf{x}) := \sum_{i\in\Omega} \mathbb{1}\{y_i = 1\} \log(f(\kappa_i(\mathbf{a}_i^T\mathbf{x} - \tau_i))) + \mathbb{1}\{y_i = -1\} \log(1 - f(\kappa_i(\mathbf{a}_i^T\mathbf{x} - \tau_i)))$$

We propose estimating $\mathbf{x}$ using the constrained maximum likelihood estimator

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} F_{\Omega,y}(\mathbf{x}) \text{ such that } \|\boldsymbol{\kappa}_\Omega \circ (\mathbf{A}_\Omega\mathbf{x} - \boldsymbol{\tau}_\Omega)\|_\infty \leq \alpha. \tag{5.8}$$

Other estimators are possible, but this one is easy to solve because it is convex and will be particularly convenient to analyze. Here, parameter α is known to the estimator and functions as some prior knowledge of where the signal lies. It is straightforward to see that if α is allowed to become arbitrarily large and $\mathbf{x}$ happens to lie very far from the origin, all else equal, this model degenerates into a *noiseless* binary sensing problem. In this case, although it would be possible to recover the direction $\mathbf{x}/\|\mathbf{x}\|$, recovery of the magnitude $\mathbf{x}$ will become impossible. Since we are interested in uniform bounds on $\|\mathbf{x} - \hat{\mathbf{x}}\|$, this will not work. Imposing the α constraint is a natural assumption and will also be important for adaptivity.

Analysis.— Taking a second-order Taylor series expansion around the ground-truth $\mathbf{x}^*$, for any $\mathbf{x}$ we have

$$F_{\Omega,y}(\mathbf{x}) = F_{\Omega,y}(\mathbf{x}^*) + \langle \nabla_x F_{\Omega,y}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle + \frac{1}{2}\langle \mathbf{x} - \mathbf{x}^*, \nabla^2_{xx} F_{\Omega,y}(\tilde{\mathbf{x}})(\mathbf{x} - \mathbf{x}^*)\rangle,$$

for some $\tilde{\mathbf{x}}$ on the line segment connecting $\mathbf{x}$ and $\mathbf{x}^*$. Our strategy will be as follows; since $F_{\Omega,y}(\hat{\mathbf{x}})$ is a maximum due to our estimator, $0 \leq F_{\Omega,y}(\hat{\mathbf{x}}) - F_{\Omega,y}(\mathbf{x}^*)$. Rearranging, we see

$$\begin{aligned}
0 &\leq \langle \nabla_x F_{\Omega,y}(x^*), \hat{\mathbf{x}} - \mathbf{x}^* \rangle + \frac{1}{2}\langle \hat{\mathbf{x}} - \mathbf{x}^*, \nabla^2_{xx} F_{\Omega,y}(\tilde{\mathbf{x}})(\hat{\mathbf{x}} - \mathbf{x}^*)\rangle \\
&\implies \langle \hat{\mathbf{x}} - \mathbf{x}^*, -\nabla^2_{xx} F_{\Omega,y}(\tilde{\mathbf{x}})(\hat{\mathbf{x}} - \mathbf{x}^*)\rangle \leq 2\langle \nabla_x F_{\Omega,y}(\mathbf{x}^*), \hat{\mathbf{x}} - \mathbf{x}^* \rangle \\
&\implies \lambda_{\min}(-\nabla^2_{xx} F_{\Omega,y}(\tilde{\mathbf{x}}))\|\hat{\mathbf{x}} - \mathbf{x}^*\| \leq 2\|\nabla_x F_{\Omega,y}(\mathbf{x}^*)\|.
\end{aligned}$$

Thus, we have

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\| \leq 2\|\nabla_x F_{\Omega,y}(\mathbf{x}^*)\| / \lambda_{\min}(-\nabla^2_{xx} F_{\Omega,y}(\tilde{\mathbf{x}})),$$

still subject to the constraints

$$\|\boldsymbol{\kappa}_\Omega \circ (\mathbf{A}_\Omega \mathbf{x}^* - \boldsymbol{\tau}_\Omega)\|_\infty \leq \alpha \quad \text{and} \quad \|\boldsymbol{\kappa}_\Omega \circ (\mathbf{A}_\Omega \hat{\mathbf{x}} - \boldsymbol{\tau}_\Omega)\|_\infty \leq \alpha.$$

In many cases, $\nabla^2_{xx} F_{\Omega,y}(\cdot)$ is negative definite so the smallest eigenvalue of $-\nabla^2_{xx} F_{\Omega,y}(\cdot)$ is well-defined and positive. This shows that the behavior of the smallest eigenvalue of the objective at points in space will become particularly important to our recovery bound and is a focus of investigation.

Details for the logistic model.— We now go through a full derivation in the logistic noise case. Write $f_i(\cdot) := f(\kappa_i(\cdot))$ for short. Note that

$$\frac{\partial F_{\Omega,y}(\mathbf{x}^*)}{\partial \mathbf{x}^*_j} = \sum_{i\in\Omega} a_{ij} \frac{f_i'(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)}{f_i(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)} \mathbb{1}\{y_i = 1\} + a_{ij} \frac{f_i'(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)}{1 - f_i(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)} \mathbb{1}\{y_i = -1\} =: \sum_{i\in\Omega} Z_{ij}$$

and thus

$$\begin{aligned}
\left\langle \nabla_x F_{\Omega,y}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \right\rangle &= \sum_j \frac{\partial F_{\Omega,y}(\mathbf{x}^*)}{\partial \mathbf{x}^*_j} (\mathbf{x}_j - \mathbf{x}^*_j) \\
&= \sum_{i\in\Omega} \sum_j a_{ij}(\mathbf{x}_j - \mathbf{x}^*_j) \frac{f_i'(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)}{f_i(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)} \mathbb{1}\{y_i = 1\} \\
&\qquad + a_{ij}(\mathbf{x}_j - \mathbf{x}^*_j) \frac{-f_i'(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)}{1 - f_i(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)} \mathbb{1}\{y_i = -1\} \\
&= \sum_{i\in\Omega} \Big(\sum_j a_{ij}(\mathbf{x}^*_j - \mathbf{x}_j) \Big) \left[\frac{f_i'(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)}{f_i(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)} \mathbb{1}\{y_i = 1\} \right. \\
&\qquad \left. - a_{ij}(\mathbf{x}^*_j - \mathbf{x}_j) \frac{f_i'(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)}{1 - f_i(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)} \mathbb{1}\{y_i = -1\} \right] =: \sum_{i\in\Omega} Z_i
\end{aligned}$$

Observe that since $\mathbb{P}\{y_i = 1\} = f_i(\mathbf{a}_i^T\mathbf{x}^* - \tau_i)$, $\mathbb{E}_y Z_i = 0$ and

$$\begin{aligned}
|Z_{ij}| &\leq \max \left\{ \left| a_i^T(\mathbf{x} - \mathbf{x}^*) \frac{f_i'(\mathbf{a}_i^T\mathbf{x} - \tau_i)}{f_i(\mathbf{a}_i^T\mathbf{x} - \tau_i)} \right|, \left| a_i^T(\mathbf{x} - \mathbf{x}^*) \frac{f_i'(\mathbf{a}_i^T\mathbf{x} - \tau_i)}{1 - f_i(\mathbf{a}_i^T\mathbf{x} - \tau_i)} \right| \right\} \\
&\leq \max_i \kappa_i |\mathbf{a}_i^T(\mathbf{x} - \mathbf{x}^*)| L_\alpha,
\end{aligned}$$

where we have defined L_α to satisfy

$$L_\alpha \geq \sup_{|\bar{\alpha}| \leq \alpha} \left\{ \frac{|f'(\bar{\alpha})|}{\min\{f(\bar{\alpha}), 1 - f(\bar{\alpha})\}} \right\}.$$

Hence, we have

$$\sum_i Z_i \leq t$$

with probability at least

$$1 - 2\exp(-t^2/(2|\Omega| L_\alpha^2 \max_i \kappa_i |\mathbf{a}_i^T(\mathbf{x}^* - \mathbf{x})|^2)).$$

Put another way, this implies that with probability at least $1 - \eta$,

$$|\langle \nabla_x F_{\Omega,y}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle| \leq L_\alpha \max_i \kappa_i |\mathbf{a}_i^T(\mathbf{x}^* - \mathbf{x})| \sqrt{2|\Omega| \log(2/\eta)}.$$

Meanwhile, letting $\zeta_i = \mathbf{a}_i^T \mathbf{x} - \tau_i$,

$$\begin{aligned} \frac{\partial^2 F_{\Omega,y}(\mathbf{x})}{\partial \mathbf{x}_j \partial \mathbf{x}_k} = &- \sum_{i \in \Omega} a_{ij} a_{ik} \frac{f_i(\zeta_i) f_i''(\zeta_i) - f_i'(\zeta_i) f_i'(\zeta_i)}{f^2(\zeta_i)} \mathbb{1}\{y_i = 1\} \\ &- a_{ij} a_{ik} \frac{(1 - f_i(\zeta_i)) f_i''(\zeta_i) + f_i'(\zeta_i) f_i'(\zeta_i)}{(1 - f_i(\zeta_i))^2} \mathbb{1}\{y_i = -1\} \end{aligned}$$

Now suppose there is a $\gamma_\alpha > 0$ such that

$$\begin{aligned} \gamma_\alpha \leq \min\Big\{ &\inf_{|\alpha_1| \leq \alpha} \frac{f'(\alpha_1) f'(\alpha_1) - f(\alpha_1) f''(\alpha_1)}{f^2(\alpha_1)}, \\ &\inf_{|\alpha_1| \leq \alpha} \frac{f'(\alpha_1) f'(\alpha_1) + (1 - f(\alpha_1)) f''(\alpha_1)}{(1 - f(\alpha_1))^2} \Big\}. \end{aligned}$$

For convenience, let $\bar{\mathbf{a}}_i := \kappa_i \mathbf{a}_i$, $\bar{\mathbf{A}}_{\Omega j} := \boldsymbol{\kappa}_\Omega \circ \mathbf{A}_{\Omega j}$, and $\bar{\mathbf{A}}_\Omega := \mathrm{diag}(\boldsymbol{\kappa}_\Omega) \mathbf{A}_\Omega$ be the measurements scaled by their respective fidelity parameters. Then we have,

$$\left[\nabla_{xx}^2 F_{\Omega,y}(\mathbf{x})\right]_{jk} \geq \left[\gamma_\alpha \sum_{i \in \Omega} \kappa_i^2 \mathbf{a}_i \mathbf{a}_i^T\right]_{jk} = \gamma_\alpha \bar{\mathbf{A}}_{\Omega j}^T \bar{\mathbf{A}}_{\Omega k}.$$

Thus, letting $\mathbf{w} := \mathbf{x} - \mathbf{x}^*$,

$$\langle \mathbf{w}, \nabla^2_{xx} F_{\Omega,y}(\tilde{\mathbf{x}})\mathbf{w}\rangle = \sum_{j,k}\Big[\frac{\partial^2 F_{\Omega,y}(\tilde{\mathbf{x}})}{\partial \mathbf{x}_j \partial \mathbf{x}_k} w_j w_k\Big] \geq \sum_{j,k} \gamma_\alpha \bar{\mathbf{A}}^T_{\Omega j}\bar{\mathbf{A}}_{\Omega k} w_j w_k = \langle \mathbf{w}, \bar{\mathbf{A}}^T_\Omega \bar{\mathbf{A}}_\Omega \mathbf{w}\rangle$$
$$\geq \gamma_\alpha \lambda_{\min}(\bar{\mathbf{A}}^T_\Omega \bar{\mathbf{A}}_\Omega)\|\mathbf{x} - \mathbf{x}^*\|^2.$$

Going back to the Taylor expansion, we can write

$$F_{\Omega,y}(\mathbf{x}) \geq F_{\Omega,y}(\mathbf{x}^*) - L_\alpha \max_i |\bar{\mathbf{a}}_i^T(\mathbf{x}^* - \mathbf{x})|\sqrt{2|\Omega|\log(2/\eta)} + \gamma_\alpha \lambda_{\min}(\bar{\mathbf{A}}^T_\Omega \bar{\mathbf{A}}_\Omega)\|\mathbf{x} - \mathbf{x}^*\|^2.$$

Since $\hat{\mathbf{x}}$ is by definition a maximizer of $F_{\Omega,y}(\cdot)$,

$$0 \leq F_{\Omega,y}(\hat{\mathbf{x}}) - F_{\Omega,y}(\mathbf{x}^*).$$

Immediately we have the following result.

Proposition 5.4.1. *Suppose measurements* $\mathbf{y}$ *are taken according to the model* (5.7) *and let* $\bar{\mathbf{a}}_i := \kappa_i \mathbf{a}_i$, $\bar{\mathbf{A}}_\Omega := \mathrm{diag}(\boldsymbol{\kappa}_\Omega)\mathbf{A}_\Omega$, *and* $\eta > 0$. *The solution* $\hat{\mathbf{x}}_{MLE}$ *to the maximum likelihood estimator* (5.8) *satisfies*

$$\begin{aligned}\|\hat{\mathbf{x}}_{MLE} - \mathbf{x}^*\| &\leq \frac{L_\alpha}{\gamma_\alpha}\frac{\max_{i\in\Omega}|\bar{\mathbf{a}}_i^T(\mathbf{x} - \mathbf{x}^*)|\sqrt{2|\Omega|\log(2/\eta)}}{\|\mathbf{x} - \mathbf{x}^*\|\lambda_{min}(\bar{\mathbf{A}}^T_\Omega \bar{\mathbf{A}}_\Omega)} \\ &\leq \frac{L_\alpha}{\gamma_\alpha}\frac{\max_i \|\bar{\mathbf{a}}_i\|\sqrt{2|\Omega|\log(2/\eta)}}{\lambda_{min}(\bar{\mathbf{A}}^T_\Omega \bar{\mathbf{A}}_\Omega)} \end{aligned} \tag{5.9}$$

with probability at least $1 - \eta$.

Initially the upper bound (5.9) appears to *grow* with increasing $|\Omega|$, contrary to intuition. However, to understand the behavior more accurately consider the case where the outer products of the item pair difference vectors (each scaled by κ_i) are on average a scaled identity, i.e., when

$$\frac{1}{|\Omega|}\sum_{i\in\Omega}\kappa_i^2 \mathbf{a}_i\mathbf{a}_i^T = \frac{1}{|\Omega|}\bar{\mathbf{A}}^T_\Omega \bar{\mathbf{A}}_\Omega = c\mathbf{I}_{k\times k}.$$

for some $c > 0$. Since

$$kc = \mathrm{Tr}\Big(\frac{1}{|\Omega|}\sum_{i\in\Omega}\kappa_i^2\mathbf{a}_i\mathbf{a}_i^T\Big) = \frac{1}{|\Omega|}\sum_{i\in\Omega}\kappa_i^2\,\mathrm{Tr}(\mathbf{a}_i\mathbf{a}_i^T) = \frac{1}{|\Omega|}\sum_{i\in\Omega}\kappa_i^2\|\mathbf{a}_i\|^2,$$

assuming that $\kappa_i = \kappa$ and $\|a_i\| = R_{\text{item}}$ for all i, we would have $c = \kappa^2 R_{\text{item}}^2/k$. Thus,

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\| \le \frac{k}{\kappa R_{\text{item}}}\frac{L_\alpha}{\gamma_\alpha}\sqrt{\frac{2\log(2/\eta)}{|\Omega|}}.$$

As expected, this decreases with an increasing number of comparisons and shows the MLE is consistent as $|\Omega| \to \infty$. If instead, we set $\kappa_i = \kappa/\|\mathbf{a}_i\|$, we would have

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\| \le \frac{k}{\kappa}\frac{L_\alpha}{\gamma_\alpha}\sqrt{\frac{2\log(2/\eta)}{|\Omega|}}.$$

By definition, the ratio L_α/γ_α grows with increasing α. In the case where $1 - f(\phi) = f(-\phi)$ and under sufficient differentiability assumptions on f, we can write,

$$\frac{L_\alpha}{\gamma_\alpha} \le \frac{f'(\alpha)}{1-f(\alpha)}\frac{f^2(\alpha)}{f'(\alpha)f'(\alpha) - f(\alpha)f''(\alpha)} = \frac{1}{f'(\alpha)/f(-\alpha) - f''(-\alpha)/f'(\alpha)}.$$

For the logistic model in particular, where

$$f(\phi) = \frac{1}{1+\exp(-\phi)}, \tag{5.10}$$

it can be shown that

$$\frac{L_\alpha}{\gamma_\alpha} \le 1 + \exp\alpha' \quad \forall\alpha \le \alpha'.$$

Corollary 5.4.2 (to Proposition 5.4.1)**.** *Fix $\Omega \subset \bar{\Omega}$ and suppose measurements $\mathbf{y}$ are taken according to model* (5.7) *in the logistic noise model* (5.10) *and let $\bar{\mathbf{a}}_i := \kappa_i\mathbf{a}_i$, $\bar{\mathbf{A}}_\Omega := \mathrm{diag}(\boldsymbol{\kappa}_\Omega)\mathbf{A}_\Omega$, and $\eta > 0$. The solution $\hat{\mathbf{x}}_{MLE}$ to the maximum likelihood estimator* (5.8) *with constraint $\alpha > 0$ satisfies*

$$\|\hat{\mathbf{x}}_{MLE} - \mathbf{x}^*\| \le (1+\exp\alpha)\frac{\max_i\|\bar{\mathbf{a}}_i\|\sqrt{2|\Omega|\log(2/\eta)}}{\lambda_{min}(\bar{\mathbf{A}}_\Omega^T\bar{\mathbf{A}}_\Omega)} \tag{5.11}$$

with probability at least $1-\eta$.

5.4.2 Measurement sets

We will relate estimation error to certain combinatorial properties of subsets of the ground set $\overline{\Omega}$. First we introduce the appropriate quantities. For any $\mathbf{x} \in \mathbb{R}^n$, define

$$\Omega_{\mathbf{x},\alpha} := \{i \in \overline{\Omega} \mid \kappa_i |\mathbf{a}_i^T \mathbf{x} - \tau_i| \leq \alpha\}.$$

and let $h_{\mathbf{x}}(\alpha) := |\Omega_{\mathbf{x},\alpha}|$. Intuitively, $h_{\mathbf{x}}(\alpha)$ is the number of measurements passing "within" α of a given point $\mathbf{x}$ (since $\|\mathbf{a}_i\|$ is not necessarily one, this must be treated as a *scaled* distance). We will suppose that for $|\overline{\Omega}|$ chosen large enough, for all $\mathbf{x}$, there should be a large enough fraction of measurements within any $\alpha \geq \alpha_0$ for some $\alpha_0 > 0$, i.e.,

$$h_{\mathbf{x}}(\alpha) \geq h'(\alpha) \quad \forall \|\mathbf{x}\| \leq R, \alpha \geq \alpha_0.$$

The purpose of α_0 is a "smallest resolution" beyond which there are not enough measurements in a given neighborhood to allow good estimation. We suppose that

$$|\Omega_{\mathbf{x},\alpha}|^{-1} \lambda_{\min}(\bar{\mathbf{A}}_{\Omega_{\mathbf{x},\alpha}}^T \bar{\mathbf{A}}_{\Omega_{\mathbf{x},\alpha}}) = \lambda_{\mathbf{x}}(\alpha) \geq \lambda'(\alpha) \quad \forall \|\mathbf{x}\| \leq R, \alpha \geq \alpha_0.$$

Together these two properties are equivalent to assuming that the set of measurements in a neighborhood of any particular point $\mathbf{x}$ is large enough and is well-conditioned. In principle, these conditions could be verified for a fixed $\overline{\Omega}$ for every $\mathbf{x}$, although this would generally intractably difficult. One may, for instance, impose an example probabilistic model for item generation $(\mathbf{p}_i, \mathbf{q}_i)$ and show that with high probability $\overline{\Omega}$ satisfies these requirements.

5.4.3 Adaptivity

Given a potentially large ground set $\overline{\Omega}$, we would like to pick a small subset Ω which still gives good sensing performance according to our upper bound (5.9). Imagine we knew of an estimate $\mathbf{x}_0 \in \mathbb{R}^k$ and a $c \geq 0$ such that $\|\mathbf{x}^* - \mathbf{x}_0\| \leq c$. Then we might seek to use this prior information to choose a small subset $\Omega \subset \overline{\Omega}$ which is well-conditioned and has small α to reduce the ratio L_α/γ_α in our upper bound.

Ideally we would choose a set of the form $\Omega \subset \Omega_{\mathbf{x}^*,\alpha}$, but $\mathbf{x}^*$ is unknown. Let $\alpha > 0$. By

the triangle inequality, for any $\Omega \subset \Omega_{x_0,\alpha}$,

$$\begin{aligned}\|\boldsymbol{\kappa}_\Omega \circ (\mathbf{A}_\Omega \mathbf{x}^* - \boldsymbol{\tau}_\Omega)\|_\infty &= \|\boldsymbol{\kappa}_\Omega \circ (\mathbf{A}_\Omega(\mathbf{x}^* - \mathbf{x}_0 + \mathbf{x}_0) - \boldsymbol{\tau}_\Omega)\|_\infty \\ &\leq \|\boldsymbol{\kappa} \circ (\mathbf{A}_\Omega \mathbf{x} - \boldsymbol{\tau}_\Omega)\|_\infty + \|\boldsymbol{\kappa} \circ \mathbf{A}_\Omega(\mathbf{x}^* - \mathbf{x}_0)\|_\infty \\ &\leq \alpha + c \max_{i \in \Omega} \kappa_i \|\mathbf{a}_i\| := \alpha'.\end{aligned}$$

This suggests that we may set $\Omega = \Omega_{x_0,\alpha}$ and use the maximum likelihood estimator with the constraint $\|\mathbf{A}_\Omega \mathbf{x} - \boldsymbol{\tau}_\Omega\| \leq \alpha'$. The estimate $\hat{\mathbf{x}}$ then satisfies

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\| \leq \frac{(1 + \exp \alpha') \max_{i \in \Omega} \kappa_i \|\mathbf{a}_i\| \sqrt{2 \log(2/\eta)}}{\lambda_{\mathbf{x}_0}(\alpha) \sqrt{h_{\mathbf{x}_0}(\alpha)}} \tag{5.12}$$

with probability at least $1 - \eta$.

In general, there may be some trade-off between the number of measurements $h_{\mathbf{x}_0}(\alpha)$ and normalized eigenvalue $\lambda_{\mathbf{x}_0}(\alpha)$ controlled by parameter α. However we will just focus on taking a subset of $\Omega_{\mathbf{x}_0,\alpha}$. This set may be much larger than necessary to reduce the error according to our theory. The solution to (5.3) described in Section 5.2 implies that if $m \geq 5k/\epsilon^2$, we can choose a subset $\Omega \subset \Omega_{\mathbf{x}_0,\alpha}$ satisfying

$$|\Omega| = m \text{ and } \lambda_{\min}(\bar{\mathbf{A}}_\Omega^T \bar{\mathbf{A}}_\Omega) \geq (1 - 3\epsilon) \sup_{\Omega' \subset \Omega_{\mathbf{x}_0,\alpha} : |\Omega'| = m} \lambda_{\min}(\mathbf{A}_{\Omega'}^T \mathbf{A}_{\Omega'}).$$

This implies

$$\frac{|\Omega|}{\lambda_{\min}(\bar{\mathbf{A}}_\Omega^T \bar{\mathbf{A}}_\Omega)} \leq (1 + 6\epsilon) \inf_{\Omega' \subset \Omega_{\mathbf{x}_0,\alpha} : |\Omega'| = m} \frac{|\Omega'|}{\lambda_{\min}(\bar{\mathbf{A}}_{\Omega'}^T \bar{\mathbf{A}}_{\Omega'})}.$$

Of course, this will only be possible if $h_{\mathbf{x}_0}(\alpha) = |\Omega_{\mathbf{x}_0,\alpha}| \geq 5k/\epsilon^2$. Combining this with Corollary 5.4.2 we have the following result.

Proposition 5.4.3. *Fix $\alpha, \epsilon > 0$ and suppose $m \geq 5k/\epsilon$ measurements $\mathbf{y}$ are taken according to the model* (5.7) *in the logistic noise setting* (5.10)*. Let $\bar{\mathbf{a}}_i := \kappa_i \mathbf{a}_i$, $\bar{\mathbf{A}}_\Omega := \operatorname{diag}(\boldsymbol{\kappa}_\Omega)\mathbf{A}_\Omega$, and $\eta > 0$. There is a polynomial-time method to choose $\Omega \subset \Omega_{0,\alpha}$ such that the solution*

$\hat{\mathbf{x}}_{MLE}$ *to the maximum likelihood estimator* (5.8) *with constraint* α *satisfies*

$$\|\hat{\mathbf{x}}_{MLE} - \mathbf{x}^*\| \leq C_{\alpha,\epsilon} \inf_{\Omega \subset \Omega_{0,\alpha}} \frac{\max_{i\in\Omega} \|\bar{\mathbf{a}}_i\| \sqrt{2m \log(2/\eta)}}{\lambda_{min}(\bar{\mathbf{A}}_\Omega^T \bar{\mathbf{A}}_\Omega)}$$

with probability at least $1 - \eta$*, where* $C_{\alpha,\epsilon} = (1 + \exp \alpha)(1 + 6\epsilon)$.

Our maximum likelihood error upper bound also suggests a natural stage-wise adaptive scheme which repeatedly reduces the error by a constant factor. Given an estimate $\mathbf{x}_0$ with $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq c$, suppose that we want to find an better estimate $\hat{\mathbf{x}}$ which satisfies $\|\hat{\mathbf{x}} - \mathbf{x}^*\| \leq c/2$ with high probability. From (5.12), where $\alpha' = \alpha + c \max_{i\in\Omega} \kappa_i \|\mathbf{a}_i\|$, we have

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\| \leq \frac{(1 + \exp \alpha') \max_{i\in\Omega} \kappa_i \|\mathbf{a}_i\| \sqrt{2 \log(2/\eta)}}{\lambda_{\mathbf{x}_0}(\alpha) \sqrt{h_{\mathbf{x}_0}(\alpha)}} \tag{5.13}$$

with probability at least $1 - \eta$. Upper bounding by $c/2$, we have that the following suffices;

$$h_{x_0}(\alpha) \geq 4 \frac{(1 + \exp(\alpha + c\phi))^2}{c^2} \cdot \frac{\phi^2}{\lambda^2_{\mathbf{x}_0}(\alpha)} \cdot 2 \log(2/\eta).$$

where $\phi \geq \max_{i\in\Omega_{\mathbf{x}_0,\alpha}} \kappa_i \|\mathbf{a}_i\|$.

5.5 Minimax lower bound for paired comparisons

In this section we give a minimax lower bound on estimation error from pairwise comparisons where noise is distributed according to according to the logistic model. Recall that each measurement i of a signal $\mathbf{x} \in \mathbb{R}^k$ is given by

$$y_i \sim \text{Bernoulli}(f_\kappa(\mathbf{a}_i^T \mathbf{x} - \tau_i)) \tag{5.14}$$

where

$$f_\kappa(\phi) := f(\kappa\phi) = \frac{1}{1 + \exp(-\kappa\phi)}.$$

Note that here we only treat the case of constant noise parameter κ.

Our result holds for any $\mathbf{x}$ such that $\|\mathbf{x}\| \leq R$ and any non-adaptive collection and estimation scheme. However, we must make some assumption on the set of items. For simplicity, assume that the items are contained inside of a sphere of radius R_{item}. The main

result of this section is as follows:

Theorem 5.5.1. *Consider taking m measurements according to the model* (5.14)*, where any item* $\mathbf{p}$ *satisfies* $\|\mathbf{p}\| \leq R_{item}$*. and* $\|\mathbf{x}\| \leq R \leq R_{item}$*. Then the minimax error* M^* *satisfies*

$$M^*(\mathbf{A}, \boldsymbol{\tau}) = \inf_{\hat{\mathbf{x}}} \sup_{\mathbf{x}} \mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|^2 \geq \frac{k \log(2)}{11\kappa^2 m R_{item}^2 (1 + \cosh(4\kappa R_{item}^2))} \wedge \frac{R^2}{5}.$$

In high-dimensional settings, this can be sharpened to

$$M^*(A, \tau) \geq C \frac{k^2}{m R_{item}^2 \kappa^2 (1+\beta)(1 + \cosh(4\kappa R_{item}^2))},$$

provided m is large enough, for some C and where $\beta \to 0$ *as* $k \to \infty$*.*

The proof of this result is based on the Fano method [185] and uses ideas from [37]. A minimax lower bound involving pairwise comparisons was given in [125], but under a very different model where items have one-dimensional scores.

Proof. Consider a collection of points $\mathcal{X} = \{\mathbf{x}_j\}_j \subset \mathbb{R}^d$, initially arbitrary. Define

$$p_{\min} := 1 - f_\kappa(\sup_{\mathbf{x}_j \in \mathscr{W}} \|\mathbf{A}\mathbf{x}_j - \tau\|_\infty) \geq 1 - f_\kappa(\alpha).$$

where α is a uniform bound on $\|\mathbf{A}\mathbf{x}_j - \tau\|_\infty$. Because $f_\kappa(\phi)$ is Lipschitz with constant $\kappa/4$, for any pair $\mathbf{x}_j, \mathbf{x}_\ell \in \mathcal{X}$, we have

$$|f_\kappa(\mathbf{a}_i^T \mathbf{w}_j - \tau_i) - f_\kappa(\mathbf{a}_i^T \mathbf{w}_\ell - \tau_i)| \leq \frac{\kappa}{4} |\mathbf{a}_i^T (\mathbf{x}_j - \mathbf{x}_\ell)|.$$

Next, using the formula [186, Appendix B] where $d_b(p, q)$ is the KL divergence between two Bernoulli random variables,

$$d_b(p, q) \leq \frac{(p-q)^2}{q(1-q)}$$

and the chain rule for divergence, we obtain

$$D(P_j, P_\ell) \leq \sum_i \frac{\kappa^2}{16 p_{\min}(1-p_{\min})} [\mathbf{a}_i^T(\mathbf{x}_j - \mathbf{x}_\ell)]^2$$
$$= \frac{\kappa^2 \|\mathbf{A}(\mathbf{x}_j - \mathbf{x}_\ell)\|^2}{16 f_\kappa(\alpha)(1-f_\kappa(\alpha))}. \tag{5.15}$$

Note that for the logistic model,

$$(f_\kappa(\alpha)(1-f_\kappa(\alpha)))^{-1} = 2(1+\cosh(\kappa\alpha)).$$

How large can a particular $|\mathbf{a}_i^T\mathbf{x} - \tau_i|$ be? We may assume that the inner term is positive because the negative case is entirely symmetric by swapping $\mathbf{p}$ and $\mathbf{q}$. Re-writing this quantity this as $\|\mathbf{p}-\mathbf{x}\|^2 - \|\mathbf{q}-\mathbf{x}\|^2$, we see that to maximize it, $\mathbf{p}$ should be chosen to be maximally far from $\mathbf{x}$ while $\mathbf{q}$ is as close as possible. In fact, if $\|\mathbf{x}\| \leq R_{\text{item}}$, we may always set $\mathbf{q} = \mathbf{x}$ regardless of $\mathbf{p}$. In this case,

$$\|\mathbf{p}-\mathbf{x}\|^2 - \|\mathbf{q}-\mathbf{x}\|^2 \leq \|\mathbf{p}-\mathbf{x}\|^2 \leq (R_{\text{item}} + \|\mathbf{x}\|)^2 \leq 4R_{\text{item}}^2$$

The second inequality is tight when $\mathbf{p}$ and $\mathbf{x}$ are oppositely oriented and equal in length. If instead $\|\mathbf{x}\| > R_{\text{item}}$, $\|\mathbf{x}-\mathbf{q}\|$ may be as small as $\|\mathbf{x}\| - R_{\text{item}}$ and $\|\mathbf{x}-\mathbf{p}\|$ may be as large as $\|\mathbf{x}\| + R_{\text{item}}$. Hence,

$$\|\mathbf{p}-\mathbf{x}\|^2 - \|\mathbf{q}-\mathbf{x}\|^2 \leq \|\mathbf{p}-\mathbf{x}\|^2 \leq (\|\mathbf{x}\| + R_{\text{item}})^2 - (\|\mathbf{x}\| - R_{\text{item}})^2 = 4R_{\text{item}}\|\mathbf{x}\|.$$

Packing set 1.— Let $\mathscr{V}$ be a 1/2-packing set for the unit $\ell 2$-ball where $|\mathscr{V}| \geq 2^k$ and let $\mathscr{X} = \delta\mathscr{V}$ for some $\delta > 0$ (to be determined). Then we have for any $j \neq \ell$,

$$\|\mathbf{x}_j - \mathbf{x}_\ell\| = \delta\|\mathbf{v}_j - \mathbf{v}_\ell\| \geq \frac{\delta}{2}$$

and since the points $\mathbf{v}_j$ are inside the unit ball,

$$\|\mathbf{x}_j - \mathbf{x}_\ell\|^2 = \delta^2\|\mathbf{v}_j - \mathbf{v}_\ell\|^2 \leq 2\delta^2.$$

Since $\mathbf{a}_i = 2(\mathbf{p}_i - \mathbf{q}_i)$,

$$\|\mathbf{A}(\mathbf{x}_j - \mathbf{x}_\ell)\|^2 \le 2\|\mathbf{A}\|_F^2 \delta^2 \le 32 m R_{\text{item}}^2 \delta^2.$$

By the convexity of the KL divergence [185] and from (5.15),

$$I(\mathbf{x};\mathbf{y}) \le \frac{1}{|\mathcal{X}|^2} \sum_{j,k=1}^{|\mathcal{X}|} D(P_j, P_\ell) \le \frac{\kappa^2}{16} 2\|\mathbf{A}\|_F^2 \delta^2 (1 + \cosh(\kappa\alpha))$$

$$\le 2\kappa^2 m R_{\text{item}}^2 \delta^2 (1 + \cosh(4\kappa(\delta R_{\text{item}} \vee R_{\text{item}}^2))).$$

We will need the following standard result.

Lemma 5.5.2 (Fano's inequality)**.** *Let P_e be the error of any estimator on a discrete set $\mathcal{X}$. Then,*

$$P_e \ge 1 - \frac{I(\mathbf{x};\mathbf{y}) + \log 2}{\log |\mathcal{X}|}.$$

We will consider two cases, the first being when the following assignment of δ gives $\delta \le R \le R_{\text{item}}$ and the second case otherwise. Specifically, in the first case, we set

$$\delta^2 = \frac{k \log(2)}{4\kappa^2 m R_{\text{item}}^2 (1 + \cosh(4\kappa R_{\text{item}}^2))}. \tag{5.16}$$

This is possible whenever

$$m \ge \frac{k \log(2)}{4\kappa^2 R_{\text{item}}^2 R^2 (1 + \cosh(4\kappa R_{\text{item}}^2)}.$$

From (5.16) and Lemma 5.5.2, $P_e \ge 1 - 1/4 - 1/k$, and using $\log |\mathcal{X}| = k \log(2)$, we have

$$M^*(\mathbf{A}, \boldsymbol{\tau}) \ge \frac{\delta^2}{2} P_e \ge \frac{k \log(2)}{11\kappa^2 m R_{\text{item}}^2 (1 + \cosh(4\kappa R_{\text{item}}^2))}.$$

Otherwise, in the case where m is not large enough to set $\delta < R$ as in (5.16), we instead hold δ at the critical value R, so P_e may only increase with smaller m. Thus we have

$$M^*(\mathbf{A}, \boldsymbol{\tau}) \ge \frac{R^2}{2}\Big(1 - \frac{1}{k} - \frac{1}{4}\Big) \ge \frac{R^2}{5}.$$

Packing set 2.— In higher dimensional settings (where k is large), we can consider an alternative packing set. We construct $\mathscr{V}$ to be an exponentially large set of unit vectors which are far apart. Specifically, we consider generating each point v_j randomly on the unit sphere. By rotational symmetry, we can bound the probability that any pair j, k are correlated as [187]

$$\mathbb{P}(|\mathbf{v}_j^T \mathbf{v}_\ell| > t/\sqrt{k}) = \mathbb{P}(|\mathbf{v}_j[1]| > t/\sqrt{k}) \leq \exp(-t^2/2).$$

By a union bound over all pairs in $\mathscr{V}$,

$$\mathbb{P}\{\exists j, k \,:\, |\mathbf{v}_j^T \mathbf{v}_\ell| > t/\sqrt{k}\} \leq \frac{1}{2}|\mathscr{V}|(|\mathscr{V}| - 1)\exp(-t^2/2).$$

Thus, if $|\mathscr{V}| = \exp(t^2/4)$, we have

$$|\mathbf{v}_j^T \mathbf{v}_\ell| \leq t/\sqrt{k} \quad \forall j \neq \ell$$

with probability strictly greater than 1/2.

Now letting

$$\mathbf{V}_j := \mathbf{v}_j \mathbf{v}_j^T - \mathbf{I}/d,$$

we have $E\mathbf{V}_j = 0$, $\|\mathbf{V}_j\| \leq 1$, and

$$\rho^2 := \|\sum_j \mathbb{E}\, \mathbf{V}_j^2\| = \||\mathscr{V}|(k-1)/k^2 \mathbf{I}\| \leq |\mathscr{V}|/k.$$

Then by matrix Bernstein [188, Theorem 6.1],

$$\mathbb{P}\big(\big\|\sum \mathbf{V}_j\big\| \geq t'\big) \leq k \exp\Big(-\frac{3t'^2}{8\rho^2}\Big) \leq k \exp\Big(-\frac{3t'^2 k}{8|\mathscr{V}|}\Big) \leq k \exp\Big(-\frac{3\beta^2|\mathscr{V}|}{8k}\Big),$$

where we have set $t' = |\mathscr{V}|\beta/k$. To bound this strictly less than 1/2, we merely need

$$\beta^2 > \frac{8k}{3|\mathscr{V}|}\log 2k \geq \frac{8k \log 4k}{3\exp(t^2/4)}.$$

In this case, we have

$$\Big\|\sum_{j=1}^{|\mathscr{V}|}\mathbf{v}_j\mathbf{v}_j^T\Big\| \le |\mathscr{V}|\beta/k \implies \frac{1}{|\mathscr{V}|}\Big\|\sum_{j=1}^{|\mathscr{V}|}\mathbf{v}_j\mathbf{v}_j^T - \frac{1}{k}\mathbf{I}\Big\| \le (1+\beta)/k.$$

Setting $t = (3/4)\sqrt{k}$, we demonstrated the existence of a set of size $\exp(3k/16)$ with

$$\|\mathbf{v}_j - \mathbf{v}_\ell\|^2 = \|\mathbf{v}_j\|^2 + \|\mathbf{v}_\ell\|^2 - 2\mathbf{v}_j^T\mathbf{v}_\ell \ge 2 - 2(3/4) = 1/2.$$

Let $\mathscr{X} = \delta\mathscr{V}$. Then we can write

$$\begin{aligned}\frac{1}{|\mathscr{X}|^2}\sum_{j,k=1}^{|\mathscr{X}|}\|A(\mathbf{x}_j - \mathbf{x}_\ell)\|^2 &= \mathrm{Tr}\Big[\mathbf{A}^T\mathbf{A}\Big(\frac{1}{|\mathscr{X}|^2}\sum_{j,k=1}^{|\mathscr{X}|}(\mathbf{x}_j - \mathbf{x}_\ell)(\mathbf{x}_j - \mathbf{x}_\ell)^T\Big)\Big]\\ &\le \|\mathbf{A}\|_F^2\|Q\| \le \|\mathbf{A}\|_F^2\delta^2(1+\beta)/k.\end{aligned}$$

where

$$Q = \frac{1}{|\mathscr{X}|}\sum_{j=1}^{|\mathscr{X}|}\mathbf{x}_j\mathbf{x}_j^T.$$

Again by the convexity of the KL divergence,

$$\begin{aligned}I(\mathbf{x};\mathbf{y}) &\le \frac{1}{|\mathscr{X}|^2}\sum_{j,k=1}^{|\mathscr{X}|}D(P_j, P_\ell) \le \frac{\kappa^2}{16}\frac{2\|\mathbf{A}\|_F^2\delta^2(1+\beta)}{k}(1+\cosh(\kappa\alpha))\\ &\le \frac{\kappa^2}{8}\frac{16mR_{\text{item}}^2\delta^2(1+\beta)}{k}(1+\cosh(4\kappa(\delta R_{\text{item}} \vee R_{\text{item}}^2))).\end{aligned}$$

Assuming $\delta \le R \le R_{\text{item}}$ and setting

$$\delta^2 = \frac{3k^2/16}{4mR_{\text{item}}^2\kappa^2(1+\beta)(1+\cosh(4\kappa R_{\text{item}}^2))},$$

we have by Lemma 5.5.2, provided m is large enough,

$$M^*(\mathbf{A},\boldsymbol{\tau}) \ge \frac{\delta^2}{2}P_e \ge \frac{3k^2/16}{11mR_{\text{item}}^2\kappa^2(1+\beta)(1+\cosh(4\kappa R_{\text{item}}^2))},$$

When m is small and $\delta = R$, we have a similar result as in the previous case. □

5.6 Application to 1-bit constrained sensing

We briefly mention another model, 1-bit constrained sensing, which is similar to the logistic noise paired comparison setting except we assume a *probit* noise model in which

$$y_i = \begin{cases} 1 & \text{w.p. } f_\sigma(\mathbf{a}_i^T \mathbf{x}^* - \tau_i) \\ -1 & \text{w.p. } 1 - f_\sigma(\mathbf{a}_i^T \mathbf{x}^* - \tau_i). \end{cases}$$

for $\mathbf{x}^* \in \mathbb{R}^n$ with $\|\mathbf{x}^*\|_0 = k$ and where

$$f_\sigma(x) := \Phi(x/\sigma) = \frac{1}{2}\big(1 + \operatorname{erf}(x/\sqrt{2\sigma^2})\big).$$

This is equivalent to assuming measurements are subject to Gaussian pre-quantization noise;

$$\mathbf{y} = \operatorname{sign}(\mathbf{A}_\Omega \mathbf{x}^* - \boldsymbol{\tau}_\Omega + \mathbf{z}_\Omega), \quad \mathbf{z}_\Omega \sim \mathcal{N}(0, \sigma^2 \mathbf{I}). \tag{5.17}$$

In this case, one can show

$$\frac{L_\alpha}{\gamma_\alpha} = \left(-\alpha/\sigma + \frac{\exp(-\alpha^2/(2\sigma^2))\sqrt{2/\pi}}{\operatorname{erfc}(\alpha/\sqrt{2\sigma^2})} \right)^{-1} \leq \sqrt{\pi/2} + \alpha/\sigma.$$

Besides the noise model, other key differences between this model and the pairwise comparison model are (i) thresholds τ_i may be chosen independently from directions $\mathbf{a}_i$, since they are not based on fixed item pairs and (ii) we explicitly assume sparsity on $\mathbf{x}$.

Where Λ is the support of $\mathbf{x}$, we can equivalently write (5.17) as

$$\mathbf{y} = \operatorname{sign}(\mathbf{A}_{\Omega,\Lambda} \mathbf{x}^*_\Lambda - \boldsymbol{\tau}_\Omega + \mathbf{z}_\Omega), \quad \mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}).$$

Restricting the columns of $\mathbf{A}_\Omega$ to the support allows us to use our previous theory. Similarly to the logistic case of Corollary 5.4.2, we can obtain an estimate using maximum likelihood.

Corollary 5.6.1 (to Proposition 5.4.1)**.** *Fix $\Omega \subset \bar{\Omega}$ and suppose measurements $\mathbf{y}$ are taken according to model* (5.17) *and let $\eta > 0$. The solution $\hat{\mathbf{x}}_{MLE}$ to the maximum likelihood*

estimator (5.8) *with constraint set to* $\alpha > 0$ *satisfies*

$$\|\hat{\mathbf{x}}_{MLE} - \mathbf{x}^*\| \leq \frac{(\sqrt{\pi/2} + \alpha/\sigma) \max_i \|\mathbf{a}_i\| \sqrt{2|\Omega| \log(2/\eta)}}{\sigma \lambda_{min}(\mathbf{A}_{\Omega,\Lambda}^T \mathbf{A}_{\Omega,\Lambda})} \tag{5.18}$$

with probability at least $1 - \eta$.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This thesis primarily studied two problems which represent examples of settings in which our ability to take measurements is fundamentally constrained by physical or situational limitations. The first of these was *constrained adaptive sensing* where measurements may be drawn from fixed, potentially coherent, ensembles. We introduced this in Chapter 2 where we show that practical improvements are possible. We returned to this problem in Chapter 5 where some partial progress to understanding the measurement selection problem is given. Although we were motivated by sparse recovery, we did not fully address the case where the support is totally unknown. Thus, while this doesn't completely solve the problem of adaptivity it does answer important questions which further the state of our understanding for a variety of problems.

Adaptivity makes little sense in situations where acquiring a measurement is cheap. Adding processing to the sensing loop would perhaps offer a theoretical improvement, but this benefit comes with a high computational cost. Traditional compressive sensing offers a substantial reduction in up-front sensing effort and keeps the major processing burden off-line, whereas adaptive sensing increases the complexity of the sensor considerably. On the other hand, if measuring a particular quantity is costly such as in problems involving human feedback, or comes at the expense of the ability to observe a different quantity, for example when controlling the positioning of robots in a sensing network, it seems prudent to minimize the number of samples necessary.

The second example of a constrained setting was given in *localization via paired comparisons* which aims to estimate a signal from binary measurements, formed by comparing distances from items drawn from a fixed set two at a time. This problem was motivated by, but is certainly not limited to, applications in recommender systems where one has an embedding of users and items into a Euclidean space which accurately captures user prefer-

ences. In such human-in-the-loop problems it is important to reduce the number of queries necessary. We claimed in Chapter 3 that in a restrictive model, exponential improvement in terms of the number of comparisons is possible using an adaptive sensing approach. We extended this work in Chapter 4 to a real, highly constrained, dataset and show empirically an adaptive approach greatly improves recovery.

6.1 Future work

Similar to our results in Chapter 3, active learning has characterized that an exponential improvement in terms of the number of queries necessary is available in some specific cases. However, distribution assumptions in the active learning literature have prevented this work from being directly and easily applied to the pairwise comparison case. One possibility for future work is to study active localization as an active learning problem, potentially allowing extension to more general settings.

Mathematically, the problem of pairwise comparisons this problem may be considered very similar to a "one-bit" sensing problem. Although both use binary measurements, the critical challenges between the two are (i) incorporation of sparsity (or other signal structure) and (ii) allowing the ambient dimension to become too large for previous approaches to be tractable. Perhaps surprisingly, there has been relatively little work in active learning in a sparse signal setting, though recent examples include [189, 190].

An extremely interesting avenue for future work in the problem of *one-bit constrained adaptive compressed sensing*. As mentioned, in this thesis we considered the constraint of binary measurements and the assumption of sparse signals largely separately. In the very high dimensional, sparse setting the class of possible signals is non-convex and much too large to use the methods of Chapter 4 directly. Hence, future work might study computational techniques for approximately evaluating and choosing measurements. This could be done, by constructing and maintaining a sketched or randomly projected version of the set of possible hypotheses. Similar ideas appeared in [191] and [192] which use random projections to operate in low-dimensional space while implicitly learning in the original higher

dimensional space.

Future work could also characterize the bit-depth/information trade-off by introducing a reduction of quantized (and un-quantized) sensing to the one-bit case. While higher-order quantizations may provide more bits per measurement, a well-designed adaptive sensing technique can perform bit-for-bit better than non-adaptive ones since there are more opportunities to adapt.

REFERENCES

[1] E. Arias-Castro, E. J. Candès, and M. Davenport, "On the fundamental limits of adaptive sensing," *IEEE Trans. Inform. Theory*, vol. 59, no. 1, pp. 472–481, 2013.

[2] R. M. Castro, J. Haupt, R. Nowak, and G. M. Raz, "Finding needles in noisy haystacks," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Las Vegas, NV, Apr. 2008.

[3] M. A. Iwen, "Group testing strategies for recovery of sparse signals in noise," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2009.

[4] M. L. Malloy and R. D. Nowak, "Near-optimal adaptive compressed sensing," *IEEE Trans. Inform. Theory*, vol. 60, no. 7, pp. 4001–4012, Jul. 2014.

[5] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Problems*, vol. 23, pp. 969–985, Jun. 2007.

[6] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 83–91, 2008.

[7] M. A. Davenport and M. B. Wakin, "Compressive sensing of analog signals using discrete prolate spheroidal sequences," *Appl. and Computational Harmonic Anal.*, vol. 33, no. 3, pp. 438–472, 2012.

[8] K. Yu, J. Bi, and V. Tresp, "Active learning via transductive experimental design," in *Proc. Int. Conf. Mach. Learning (ICML)*, ser. ICML '06, New York, NY: ACM, 2006.

[9] E. Novak, "On the power of adaption," *J. Complexity*, vol. 12, no. 3, pp. 199–237, 1996.

[10] E. Candès, "Compressive sampling," in *Proc. Int. Congr. of Math*, Madrid, Spain, Aug. 2006.

[11] D. Donoho, "Compressed sensing," *IEEE Trans Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[12] M. Davenport, M. Duarte, Y. Eldar, and G. Kutyniok, "Introduction to compressed sensing," in *Compressed Sensing: Theory and Applications*, Y. Eldar and G. Kutyniok, Eds., Cambridge, UK: Cambridge University Press, 2012.

[13] E. Candes and M. Wakin, "An Introduction To Compressive Sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[14] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[15] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[16] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[17] S. Chen, D. Donoho, and M. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, Jan. 1, 1998.

[18] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, Dec. 2007.

[19] R. E. Carrillo, L. F. Polanía, and K. E. Barner, "Iterative hard thresholding for compressed sensing with partially known support," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2011.

[20] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[21] Y. Pati, R. Rezaifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1993.

[22] D. Needell and R. Vershynin, "Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 2, pp. 310–316, Apr. 2010.

[23] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[24] Y. C. Eldar, "Compressed sensing of analog signals," *Submitt. IEEE Trans Signal Process.*, 2008.

[25] M. G. Moore, **A. K. Massimino**, and M. A. Davenport, "Randomized multi-pulse time-of-flight mass spectrometry," in *Proc. IEEE Int. Work. on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Cancun, Mexico, Dec. 2015.

[26] M. Lustig, D. Donoho, J. Santos, and J. Pauly, "Compressed sensing mri," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 72–82, Mar. 2008.

[27] M. F. Duarte and Y. C. Eldar, "Structured Compressed Sensing: From Theory to Applications," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4053–4085, Sep. 2011.

[28] A. S. Charles and C. J. Rozell, "Spectral superresolution of hyperspectral imagery using reweighted ℓ_1 spatial filtering," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 3, pp. 602–606, 2014.

[29] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, ser. Appl. and Numerical Harmonic Anal. New York, NY: Springer New York, 2013.

[30] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Math.*, vol. 346, pp. 589–592, 9-10 2008.

[31] R. G. Baraniuk, M. A. Davenport, R. A. DeVore, and M. B. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constr. Approx.*, vol. 28, no. 3, pp. 253–263, 2008.

[32] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin, "Certifying the restricted isometry property is hard," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3448–3450, Jun. 2013.

[33] A. Bastounis and A. C. Hansen, "On the absence of the RIP in real-world applications of compressed sensing and the RIP in levels," *arXiv.org:1411.4449*, Nov. 2014.

[34] M. A. Davenport, **A. K. Massimino**, D. Needell, and T. Woolf, "Constrained adaptive sensing," *IEEE Trans. Signal Processing*, vol. 64, no. 20, pp. 5437–5449, Oct. 2016.

[35] J. Bourgain, S. J. Dilworth, K. Ford, S. V. Konyagin, and D. Kutzarova, "Breaking the k^2 barrier for explicit RIP matrices," in *Proc. Annu. ACM Symp. on Theory of Computing*, ACM, 2011.

[36] D. G. Mixon, "Explicit matrices with the restricted isometry property: Breaking the square-root bottleneck," in *Compressed Sensing and Its Applications*, Springer, 2015, pp. 389–417.

[37] E. Candès and M. Davenport, "How well can we estimate a sparse vector?" *Appl. Comput. Harmon. Anal.*, vol. 34, no. 2, pp. 317–323, 2013.

[38] M. Elad, "Optimized projections for compressed sensing," *IEEE Transactions on Signal Processing*, vol. 55, no. 12, pp. 5695–5702, Dec. 2007.

[39] S. Ji and L. Carin, "Bayesian compressive sensing and projection optimization," in *Proceedings of the 24th International Conference on Machine Learning*, ACM, 2007.

[40] J. Haupt, R. Castro, and R. Nowak, "Adaptive discovery of sparse signals in noise," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2008.

[41] R. Castro and R. Nowak, "Active sensing and learning," *Found. Appl. Sens. Manag.*, pp. 177–200, 2009.

[42] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Selective sampling for sparse signal recovery," in *Proc. Int. Conf. Art. Intell. Stat. (AISTATS)*, Clearwater Beach, FL, Apr. 2009.

[43] J. D. Haupt, R. G. Baraniuk, R. M. Castro, and R. D. Nowak, "Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements," in *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference On*, IEEE, 2009.

[44] D. M. Malioutov, S. R. Sanghavi, and A. S. Willsky, "Sequential compressed sensing," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 2, pp. 435–444, 2010.

[45] P. Indyk, E. Price, and D. Woodruff, "On the power of adaptivity in sparse recovery," in *Proc. IEEE Symp. Found. Comp. Science (FOCS)*, Palm Springs, CA, Oct. 2011.

[46] L. Wasserman, *All of Nonparametric Statistics*, G. Casella, S. Fienberg, and I. Olkin, Eds., ser. Springer Texts in Statistics. Springer, Mon Apr 11 07:05:38 2011 EDT.

[47] R. Castro, "Adaptive sensing performance lower bounds for sparse signal detection and support estimation," *Bernoulli*, vol. 20, no. 4, pp. 2217–2246, 2014.

[48] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535, Sep. 1952.

[49] Valerii V. Fedorov and Sergei L. Leonov, *Optimal Design for Nonlinear Response Models*. 2014.

[50] A. Wald, *Sequential Analysis*. Courier Corporation, 1973.

[51] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Stat.*, vol. 16, no. 2, pp. 117–186, 1945.

[52] D. Siegmund, *Sequential Analysis*, ser. Springer Series in Statistics. New York: Springer-Verlag, 1985.

[53] H. Chernoff, *Sequential Analysis and Optimal Design*. Soc. Ind. Appl. Mathematics, Jan. 1, 1972.

[54] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, N.J.: Wiley, 2006.

[55] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback–I: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. 12, no. 2, pp. 172–182, Apr. 1966.

[56] O. Shayevitz and M. Feder, "Optimal Feedback Communication Via Posterior Matching," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1186–1222, Mar. 2011.

[57] R. Ma and T. P. Coleman, "Generalizing the posterior matching scheme to higher dimensions via optimal transportation," in *Proc. Commun., Control, and Computing (Allerton)*, IEEE, 2011.

[58] D. Braziunas, "Pomdp solution methods," *Univ. Tor. Tech Rep*, 2003.

[59] R. Zahedi, L. W. Krakow, E. K. P. Chong, and A. Pezeshki, "Adaptive compressive sampling using partially observable markov decision processes," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Mar. 2012.

[60] L. W. Krakow, R. Zahedi, E. K. P. Chong, and A. Pezeshki, "Adaptive compressive sensing in the presence of noise and erasure," in *Proc. IEEE Global Conf. on Signal and Inform. Process.*, Dec. 2013.

[61] R. Zahedi, L. W. Krakow, E. K. P. Chong, and A. Pezeshki, "Adaptive Estimation of Time-Varying Sparse Signals," *IEEE Access*, vol. 1, pp. 449–464, 2013.

[62] J.-Y. Audibert, R. Munos, and C. Szepesvari, "Exploration-exploitation trade-off using variance estimates in multi-armed bandits," *Theor. Comp Sci*, vol. 410, pp. 1876–1902, 2009.

[63] R. Ganti and A. G. Gray, "Building Bridges: Viewing Active Learning from the Multi-Armed Bandit Lens," Thu May 16 19:03:57 2013 EDT.

[64] J. N. Laska and R. G. Baraniuk, "Regime Change: Bit-Depth Versus Measurement-Rate in Compressive Sensing," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3496–3505, Jul. 2012.

[65] J. N. Laska, Z. Wen, W. Yin, and R. G. Baraniuk, "Trust, But Verify: Fast and Accurate Signal Recovery From 1-Bit Compressive Measurements," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5289–5301, Nov. 2011.

[66] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Inf. Inference*, vol. 3, no. 3, pp. 189–223, Sep. 2014.

[67] M. Slawski and P. Li, "Linear signal recovery from b-bit-quantized linear measurements: Precise analysis of the trade-off between bit depth and number of measurements," Jul. 9, 2016.

[68] L. Jacques, K. Degraux, and C. De Vleeschouwer, "Quantized Iterative Hard Thresholding: Bridging 1-bit and High-Resolution Quantized Compressed Sensing," *Int. Conf. Sampl. Theory Appl. SampTA*, May 8, 2013. arXiv: `1305.1786 [cs, math]`.

[69] L. Jacques, J. Laska, P. Boufounos, and R. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Trans. Inform. Theory*, vol. 59, no. 4, Apr. 2013.

[70] Y. Plan and R. Vershynin, "Robust 1-bit Compressed Sensing and Sparse Logistic Regression: A Convex Programming Approach," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 482–494, Jan. 2013.

[71] A. Gupta, R. Nowak, and B. Recht, "Sample complexity for 1-bit compressed sensing and sparse classification," in *2010 IEEE International Symposium on Information Theory*, 2010.

[72] K. Knudson, R. Saab, and R. Ward, "One-bit compressive sensing with norm estimation," *IEEE Trans. Inform. Theory*, vol. 62, no. 5, pp. 2748–2758, 2016.

[73] L. Jacques and V. Cambareri, "Time for dithering: Fast and quantized random embeddings via the restricted isometry property," Jul. 5, 2016.

[74] R. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters, "Exponential decay of reconstruction error from binary measurements of sparse signals," *arXiv:1407.8246*, 2014.

[75] J. Li, M. M. Naghsh, S. J. Zahabi, and M. Modarres-Hashemi, "Compressive radar sensing via one-bit sampling with time-varying thresholds," in *Proc. Asilomar Conf. on Signals, Systems and Comput.*, Nov. 2016.

[76] U. S. Kamilov, A. Bourquard, A. Amini, and M. Unser, "One-bit measurements with adaptive thresholds," *Signal Process. Lett. IEEE*, vol. 19, no. 10, pp. 607–610, 2012.

[77] V. N. Vapnik, "An Overview of Statistical Learning Theory," *IEEE Trans. Neural Netw.*, vol. 10, Sep. 1999.

[78] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, ser. Stochastic Modelling and Applied Probability. New York, NY: Springer New York, 1996, vol. 31.

[79] M. E. Ahsen and M. Vidyasagar, "A PAC learning approach to one-bit compressed sensing," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2015.

[80] M. E. Ahsen and M. Vidyasagar, "An approach to one-bit compressed sensing based on probably approximately correct learning theory," in *Proc. IEEE Conf. Decision and Control (CDC)*, Dec. 2015.

[81] S. Spencer, "Noisy 1-Bit Compressed Sensing Embeddings Enjoy a Restricted Isometry Property," Apr. 12, 2016. arXiv: `1604.03499 [cs, math]`.

[82] B. Settles, *Active Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, Jun. 30, 2012, 100 pp.

[83] S. Shalev-Shwartz and S. Ben-David, "Understanding machine learning: From theory to algorithms," 2014.

[84] Steve Hanneke, "Theory of Active Learning," Sep. 22, 2014.

[85] S. Dasgupta, "Analysis of a greedy active learning strategy," in *Advances in Neural Information Processing Systems*, 2004.

[86] M. F. Balcan and P. M. Long, "Active and passive learning of linear separators under log-concave distributions," Nov. 5, 2012. arXiv: `1211.1082 [cs, math, stat]`.

[87] A. Gonen, S. Sabato, and S. Shalev-Shwartz, "Efficient active learning of halfspaces: An aggressive approach," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2583–2615, 2013.

[88] M. Davenport and E. Arias-Castro, "Compressive binary search," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Cambridge, MA, Jul. 2012.

[89] R. D. Nowak, "The Geometry of Generalized Binary Search," Oct. 22, 2009. arXiv: `0910.4397 [cs, math, stat]`.

[90] R. Nowak, "Noisy generalized binary search," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2009.

[91] S. S. Vempala, "Recent progress and open problems in algorithmic convex geometry," 2010.

[92] B. Cousins and S. Vempala, "Gaussian cooling and $O^*(N^3)$ algorithms for volume and gaussian volume," Sep. 21, 2014. arXiv: `1409.6011 [cs, math]`.

[93] P. M. Vaidya, "A new algorithm for minimizing convex functions over convex sets," *Mathematical Programming*, vol. 73, no. 3, pp. 291–341, Jun. 1, 1996.

[94] M. A. Davenport, **A. K. Massimino**, D. Needell, and T. Woolf, "Constrained adaptive sensing," *arXiv.org:1506.05889*, Jul. 2016.

[95] M. A. Davenport, **A. K. Massimino**, D. Needell, and T. Woolf, "Constrained adaptive sensing," in *Proc. Work. on Signal Process. with Adaptive Sparse Structured Representations (SPARS)*, Cambridge, United Kingdom, Jul. 2015.

[96] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Trans. Inform. Theory*, vol. 57, no. 9, pp. 6222–6235, 2011.

[97] N. Bakhvalov, "On the optimality of linear methods for operator approximation in convex classes of functions," *USSR Computational Mathematics and Math. Physics*, vol. 11, no. 4, pp. 244–249, 1971.

[98] H. Woźniakowski, "A survey of information-based complexity," *USSR Computational Mathematics and Math. Physics*, vol. 1, no. 1, pp. 11–44, 1985.

[99] R. Castro, R. Willett, and R. Nowak, "Faster rates in regression via active learning," in *Proc. Adv. in Neural Processing Systems (NIPS)*, 2005.

[100] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[101] H. Nickisch and M. Seeger, "Compressed sensing and bayesian experimental design," in *Proc. Int. Conf. Mach. Learning (ICML)*, Helsinki, Finland, Jul. 2008.

[102] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 83–91, 2008.

[103] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Nearly optimal sparse Fourier transform," in *Proc. ACM Symp. Theory of Comput.*, New York, NY, May 2012.

[104] F. Pukelsheim, *Optimal Design of Experiments*. Soc. Ind. Appl. Mathematics, Jan. 2006.

[105] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[106] I. Daubechies, *Ten lectures on wavelets*. Philadelphia, PA: SIAM, 1992.

[107] S. Becker, E. Candès, and M. Grant, "Templates for convex cone problems with applications to sparse signal recovery," *Math. Prog. Comput.*, vol. 3, no. 3, pp. 165–218, 2011.

[108] S. Becker, E. Candès, and M. Grant, "Templates for first-order conic solvers user guide, version 1.3 release 2," Stanford Univ., Tech. Rep., 2014.

[109] M. Duarte, M. Wakin, and R. Baraniuk, "Fast reconstruction of piecewise smooth signals from random projections," in *Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS)*, Rennes, France, Nov. 2005.

[110] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using Hidden Markov Models," *IEEE Trans. Signal Processing*, vol. 46, no. 4, pp. 886–902, 1998.

[111] F. Krahmer and R. Ward, "Stable and robust sampling strategies for compressive imaging," *IEEE Trans. Image Processing*, vol. 23, no. 2, pp. 612–622, 2014.

[112] F. Krahmer, D. Needell, and R. Ward, "Compressive sensing with redundant dictionaries and structured measurements," *SIAM J. Math. Anal.*, 2015.

[113] E. van den Berg and M. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM J. Sci. Comput.*, vol. 31, no. 2, pp. 890–912, 2008.

[114] E. van den Berg and M. Friedlander, *SPGL1: A solver for large-scale sparse reconstruction*, Jun. 2007.

[115] M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf, "Optimization of k-space trajectories for compressed sensing by Bayesian experimental design," *Magnetic Resonance in Medicine*, vol. 63, no. 1, pp. 116–126, 2010.

[116] H. Nickisch, R. Pohmann, B. Schölkopf, and M. Seeger, "Bayesian experimental design of magnetic resonance imaging sequences," in *Proc. Adv. in Neural Processing Systems (NIPS)*, Vancouver, BC, Dec. 2009.

[117] M. Seeger, "Speeding up magnetic resonance image acquisition by Bayesian multislice adaptive compressed sensing," in *Proc. Adv. in Neural Processing Systems (NIPS)*, Vancouver, BC, Dec. 2009.

[118] A. Petrosian and F. Meyer, *Wavelets in signal and image analysis: From theory to practice*. Springer Science & Business Media, 2001.

[119] **A. K. Massimino** and M. A. Davenport, "As you like it: Localization via paired comparisons," *Submitted*, Feb. 2018. eprint: `1802.10489`.

[120] **A. K. Massimino** and M. A. Davenport, "Binary stable embedding via paired comparisons," in *Proc. IEEE Work. on Statistical Signal Process. (SSP)*, Palma de Mallorca, Spain, Jun. 2016.

[121] **A. K. Massimino** and M. A. Davenport, "The geometry of random paired comparisons," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, New Orleans, LA, 2017.

[122] N. Ailon, "Active learning ranking from pairwise preferences with almost optimal query complexity," in *Proc. Adv. in Neural Inform. Process. Systems (NIPS)*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2011, pp. 810–818.

[123] M. Davenport, "Lost without a compass: Nonmetric triangulation and landmark multidimensional scaling," in *Proc. IEEE Int. Work. Comput. Adv. Multi-Sensor Adaptive Process. (CAMSAP)*, Saint Martin, Dec. 2013.

[124] B. Eriksson, "Learning to top-K search using pairwise comparisons," in *Proc. of the Sixteenth Int. Conf. on Artificial Intelligence and Statistics*, 2013.

[125] N. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright, "Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence," vol. 17, no. 58, pp. 1–47, 2016.

[126] C. Coombs, "Psychological scaling without a unit of measurement," *Psych. Rev.*, vol. 57, no. 3, pp. 145–158, 1950.

[127] G. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psych. review*, vol. 63, no. 2, p. 81, 1956.

[128] H. David, *The method of paired comparisons*. London, UK: Charles Griffin & Company Limited, 1963.

[129] F. Radlinski and T. Joachims, "Active exploration for learning rankings from clickthrough data," in *Proc. ACM SIGKDD Int. Conf. on Knowledge, Discovery, and Data Mining*, San Jose, CA, Aug. 2007.

[130] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, and S. Belongie, "Generalized non-metric multidimensional scaling," in *Proc. Int. Conf. Art. Intell. Stat. (AISTATS)*, San Juan, PR, Mar. 2007.

[131] J. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proc. Int. Conf. Machine Learning (ICML)*, Bonn, Germany, Aug. 2005.

[132] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[133] B. Dubois, "Ideal point versus attribute models of brand preference: A comparison of predictive validity," *Adv. Consumer Research*, vol. 2, no. 1, pp. 321–333, 1975.

[134] A. Maydeu-Olivares and U. Böckenholt, "Modeling preference data," *The SAGE handbook of quantitative methods in psychology*, pp. 264–282, 2009.

[135] K. G. Jamieson and R. Nowak, "Active ranking using pairwise comparisons," in *Proc. Adv. in Neural Processing Systems (NIPS)*, Granada, Spain, Dec. 2011.

[136] K. Jamieson and R. Nowak, "Low-dimensional embedding using adaptively selected ordinal data," in *Proc. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, 2011.

[137] F. Wauthier, M. Jordan, and N. Jojic, "Efficient ranking from pairwise comparisons," in *Proc. Int. Conf. Machine Learning (ICML)*, Atlanta, GA, Jun. 2013.

[138] M. O'Shaughnessy and M. Davenport, "Localizing users and items from paired comparisons," in *Proc. IEEE Int. Work. on Mach. Learning for Signal Process. (MLSP)*, Vietri sul Mare, Salerno, Italy, Sep. 2016.

[139] Y. Lu and S. Negahban, "Individualized rank aggregration using nuclear norm regularization," *arXiv:1410.0860*, 2014.

[140] D. Park, J. Neeman, J. Zhang, S. Sanghavi, and I. Dhillon, "Preference completion: Large-scale collaborative ranking form pairwise comparisons," in *Proc. Int. Conf. Mach. Learning (ICML)*, Lille, France, Jul. 2015.

[141] S. Oh, K. Thekumparampil, and J. Xu, "Collaboratively learning preferences from ordinal data," in *Proc. Adv. in Neural Processing Systems (NIPS)*, Montréal, Québec, Dec. 2014.

[142] A. Brieden, P. Gritzmann, R. Kannan, V. Klee, L. Lovász, and M. Simonovits, "Deterministic and randomized polynomial-time approximation of radii," *Mathematika*, vol. 48, no. 1-2, pp. 63–105, 2001.

[143] R. Buck, "Partition of space," *Amer. Math. Monthly*, vol. 50, no. 9, pp. 541–544, 1943.

[144] M. Giaquinta and G. Modica, *Mathematical analysis: An introduction to functions of several variables*. Springer Science & Business Media, 2010, p. 105.

[145] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*. New York, NY: Wiley, 1994, vol. 1.

[146] M. Spruill, "Asymptotic distribution of coordinates on high dimensional spheres," *Elec. Comm. in Prob.*, vol. 12, pp. 234–247, 2007.

[147] L. Grenié and G. Molteni, "Inequalities for the beta function," *Math. Inequal. Appl*, vol. 18, no. 4, pp. 1427–1442, 2015.

[148] F. Qi, "Bounds for the ratio of two gamma functions," *J. Inequal. Appl.*, vol. 2010, no. 1, p. 493 058, 2010.

[149] F. Pérez-Cruz, J. Weston, D. Herrmann, and B. Schölkopf, "Extension of the ν-SVM range for classification," *NATO Science Series Sub Series III Computer and Systems Sciences*, vol. 190, pp. 179–196, 2003.

[150] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2001.

[151] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[152] L. Thurstone, "A law of comparative judgment," *Psych. Rev.*, vol. 34, no. 4, p. 273, 1927.

[153] P.-H. Chen, C.-J. Lin, and B. Schölkopf, "A tutorial on ν-support vector machines," *Appl. Stoch. Models Bus. Ind.*, vol. 21, no. 2, pp. 111–136, 2005.

[154] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*. Academic Press, 2007.

[155] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai, "Adaptively Learning the Crowd Kernel," in *International Conference on Machine Learning (ICML)*, May 2011.

[156] L. Van Der Maaten and K. Weinberger, "Stochastic triplet embedding," in *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, IEEE, 2012.

[157] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[158] S. Negahban, S. Oh, and D. Shah, "Iterative ranking from pair-wise comparisons," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012.

[159] N. B. Shah and M. J. Wainwright, "Simple, robust and optimal ranking from pairwise comparisons.," *Journal of machine learning research*, vol. 18, pp. 199–1, 2017.

[160] Y. Guo, P. Tian, J. Kalpathy-Cramer, S. Ostmo, J. P. Campbell, M. F. Chiang, D. Erdogmus, J. G. Dy, and S. Ioannidis, "Experimental design under the bradley-terry model.," in *IJCAI*, 2018.

[161] F. Bach *et al.*, "Self-concordant analysis for logistic regression," *Electronic Journal of Statistics*, vol. 4, pp. 384–414, 2010.

[162] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning*, ACM, 2007.

[163] Y. Chen and C. Suh, "Spectral mle: Top-k rank aggregation from pairwise comparisons," in *International Conference on Machine Learning*, 2015.

[164] K. G. Jamieson, S. Katariya, A. Deshpande, and R. D. Nowak, "Sparse dueling bandits.," in *AISTATS*, 2015.

[165] A. Saumard and J. A. Wellner, "Log-concavity and strong log-concavity: A review," *Statistics surveys*, vol. 8, p. 45, 2014.

[166] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[167] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014.

[168] M. Wilber, I. S. Kwak, D. Kriegman, and S. Belongie, "Learning concept embeddings with combined human-machine expertise," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[169] M. J. Wilber, I. S. Kwak, and S. J. Belongie, "Cost-effective hits for relative similarity comparisons," in *Second AAAI conference on human computation and crowdsourcing*, 2014.

[170] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of statistical software*, vol. 76, no. 1, 2017.

[171] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of markov chain monte carlo*. CRC press, 2011.

[172] C. Frowd, V. Bruce, M. Pitchford, C. Gannon, M. Robinson, C. Tredoux, J. Park, A. Mcintyre, and P. J. Hancock, "Evolving the face of a criminal: How to search a face space more effectively," *Soft Computing*, vol. 15, no. 1, pp. 61–70, 2011.

[173] E. E. Ventura, J. N. Davis, and M. I. Goran, "Sugar content of popular sweetened beverages based on objective laboratory analysis: Focus on fructose content," *Obesity*, vol. 19, no. 4, pp. 868–874, 2011.

[174] L. Lovász and S. Vempala, "The geometry of logconcave functions and sampling algorithms," *Random Structures & Algorithms*, vol. 30, no. 3, pp. 307–358, 2007.

[175] A. Marsiglietti and V. Kostina, "A lower bound on the differential entropy of log-concave random vectors with applications," *Entropy*, vol. 20, no. 3, p. 185, 2018.

[176] S. G. Bobkov and M. M. Madiman, "On the problem of reversibility of the entropy power inequality," in *Limit theorems in probability, statistics and number theory*, Springer, 2013, pp. 61–74.

[177] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2010.

[178] A. R. Klivans, P. M. Long, and A. K. Tang, "Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, Springer, 2009, pp. 588–600.

[179] Z. Allen-Zhu, Y. Li, A. Singh, and Y. Wang, "Near-Optimal Discrete Optimization for Experimental Design: A Regret Minimization Approach," *ArXiv e-prints*, vol. abs/1711.05174, Nov. 2017.

[180] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, 2011.

[181] V. Fedorov, *Theory of Optimal Experiments*. New York, NY: Academic Press, 1972.

[182] N. J. Harvey and N. Olver, "Pipage rounding, pessimistic estimators and matrix concentration," in *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics, 2014.

[183] A. Agresti, *Categorical data analysis*. John Wiley & Sons, 2003, vol. 482.

[184] T. Hastie and R. Tibshirani, "[Generalized additive models]: Rejoinder," *Statist. Sci.*, vol. 1, no. 3, pp. 314–318, Aug. 1986.

[185] S. Verdu *et al.*, "Generalizing the fano inequality," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1247–1251, 1994.

[186] S. Gerchinovitz and T. Lattimore, "Refined lower bounds for adversarial bandits," in *Advances in Neural Information Processing Systems*, 2016.

[187] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[188] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[189] L. Zhang, J. Yi, and R. Jin, "Efficient algorithms for robust one-bit compressive sensing," in *Proc. Int. Conf. Mach. Learning (ICML)*, Beijing, China, 2014.

[190] P. Awasthi, M.-F. Balcan, N. Haghtalab, and H. Zhang, "Learning and 1-bit Compressed Sensing under Asymmetric Noise," in *Conference on Learning Theory (COLT)*, 2016.

[191] M.-F. Balcan, A. Blum, and S. Vempala, "Kernels as features: On kernels, margins, and low-dimensional mappings," *Mach Learn*, vol. 65, no. 1, pp. 79–94, Oct. 1, 2006.

[192] R. I. Arriaga and S. Vempala, "An algorithmic theory of learning: Robust concepts and random projection," *Mach Learn*, vol. 63, no. 2, pp. 161–182, May 1, 2006.